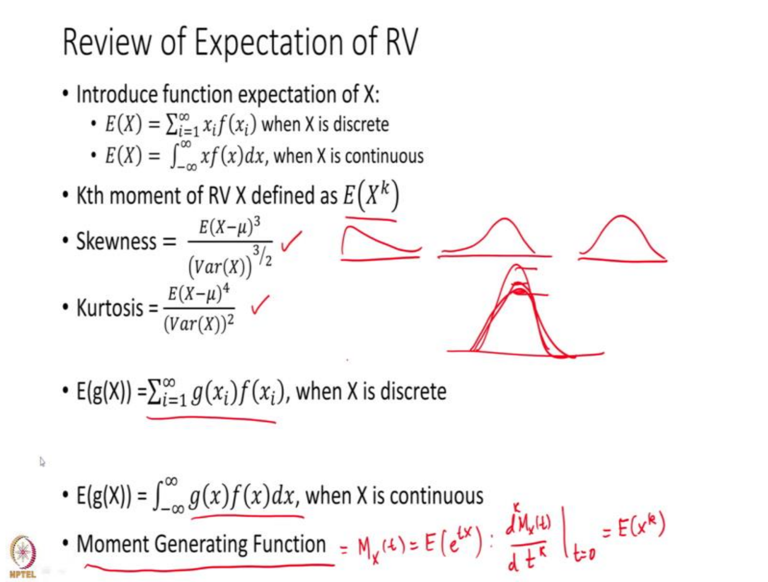
Dealing with Materials Data:Collection, Analysis and Interpretation

Prof. M P Gururajan.
Prof. Hina A.Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay

Lecture 6
Random variable and Expectation III

(Refer Slide Time: 00:38)



Hello and welcome to the course on Dealing with Materials Data. We are from the previous two sessions we are introducing here the aspect of random variable and its expectation. Let us continue first we will review what we have done so far, we introduced the function of expectation of a random variable of X as in the case of discrete random variable.

We define expected value of X, when X is discrete

$$E(X) = \sum_{i=1}^{\infty} x_i f(x_i)$$

The probability mass function of X in the case of a continues random variable We define it as

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

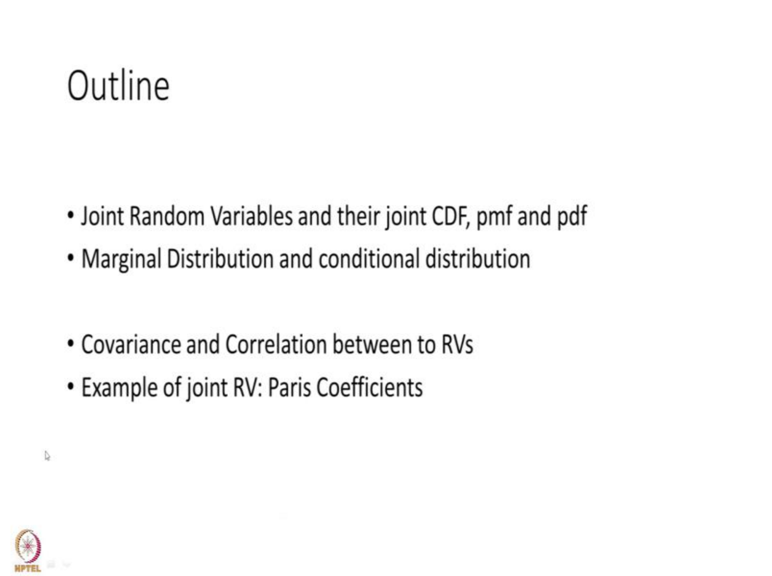
We also define the Kth moment of X as expected value of X to the power K.and we define a moment generating function which is defined as M of t and this moment generating function has a property that, let us write it down quickly here, I have mentioned it in the bottom, that this moment generating function M sub x of t is defined as an expected value of exponential to the power tx and the property of that is that the derivative of M of x of t with respect to t dt.

If you take the Kth that is K to the power this to the power this at the value of t is equal to 0, it gives you the expected value of X to the power K. So that is why it is called Kth it is called the moment generating function, so we defined the Kth moment Kth raw moment of expected random variable X and we also define the moment generating function of the random variable X along with that we define the two important measures. One is called the measure of the coefficient of skewness and the coefficient of kurtosis, just recall that skewness defines whether your function is positively skewed or is it negatively skewed or is it symmetric.

In the case of kurtosis it actually tries to see whether the tip is sharper than the normal or it is less than the normal, while the normal will be like something like this. So, it defines if it is sharper then the kurtosis will be more than 3. If it is flatter it will be less than 3. And if it is like a normal curve absolutely symmetric bell shaped normal curve it will be 3.

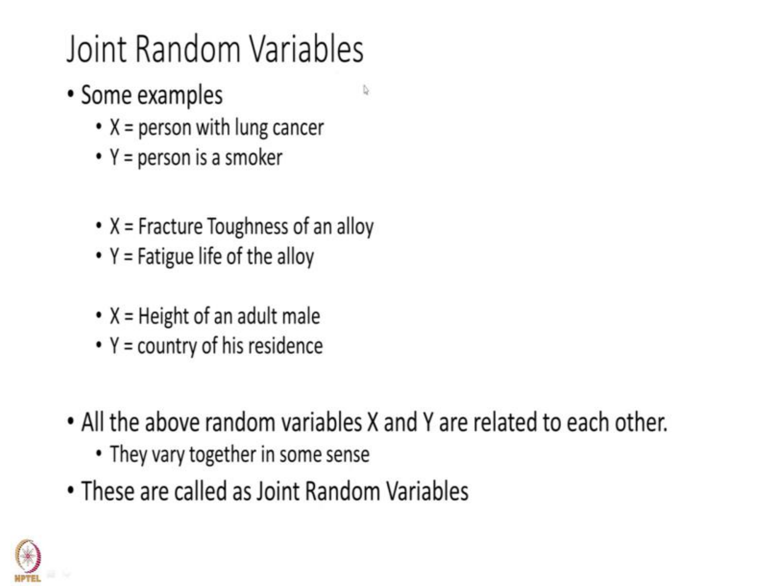
So, we also measured introduced those measures, we also introduced expected value of a function of a random variable X as in the case of discrete, it is a summation i is equal to 1 to infinity g(xi) multiplied by probability mass function of xi and in the case of continues random variable it is integral minus infinity to infinity gx multiplied by probability density function of X dx.

(Refer Slide Time: 04:14)



So, in this particular session we would like to introduce what is known as joint random variables and their joint cumulative distribution function, probability mass function, in case they are discrete and probability density function in case it is continues. Then we will introduce marginal distribution function and the conditional distribution. We will also define the co-variance and correlation between such two random variables and finally we will give an example in terms of Paris coefficients, estimated Paris coefficients.

(Refer Slide Time: 04:56)

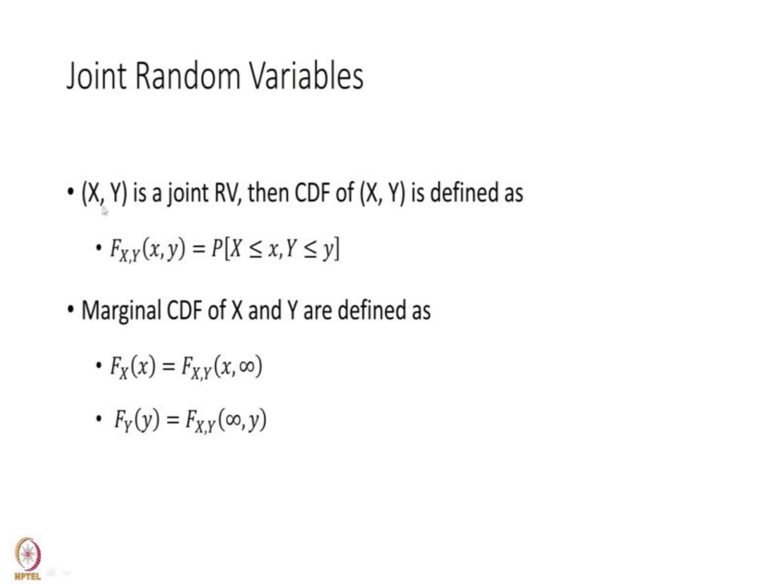


Let us move on, the joint random variables occur very naturally in our day to day life. For example I have given a few examples here, if a person has a lung cancer and a person is a smoker, these two are correlated events and they vary together. Similarly in the case of metallurgy and material science world the fracture toughness of an alloy and the fatigue life of an alloy are also closely related and they seem to vary together in some sense.

Similarly a height of an adult male or a female and the country of his residence also has an effect and therefore they also some kind of vary together. So, all of the above random variables we find that though they are two different random variables they do vary together in certain sense. Such variables are called joint random variables.

Right now we are going to discuss about 2 joint random variables but we should remember that there is no it is not necessary that we have only 2 joint random variables, we may have more for example you may have 3 joint random variables, you may have 4 joint random variables, but the theory is going to be more or less the same. So, we are going to start off with a 2 random variable, jointly distributed random variables and be the other case will be left to you for understanding because it is a simple generalization.

(Refer Slide Time: 06:50)



So, let us define in we write it as in a bracket X and Y to show the jointness between the two. So XY is a joint random variables. Then the cumulative distribution function, remember I had said earlier also with any random variables there is one entity always attached to it and that is called cumulative distribution function. Whether it is continues or it is discrete, this quantity is always attached to it and remembers it actually comes from the definition of random variable itself because it comes from the probability space.

So, here we define the cumulative distribution function of XY as

$$F_{X,Y}(x,y) = P[X \leq x, Y \leq y]$$

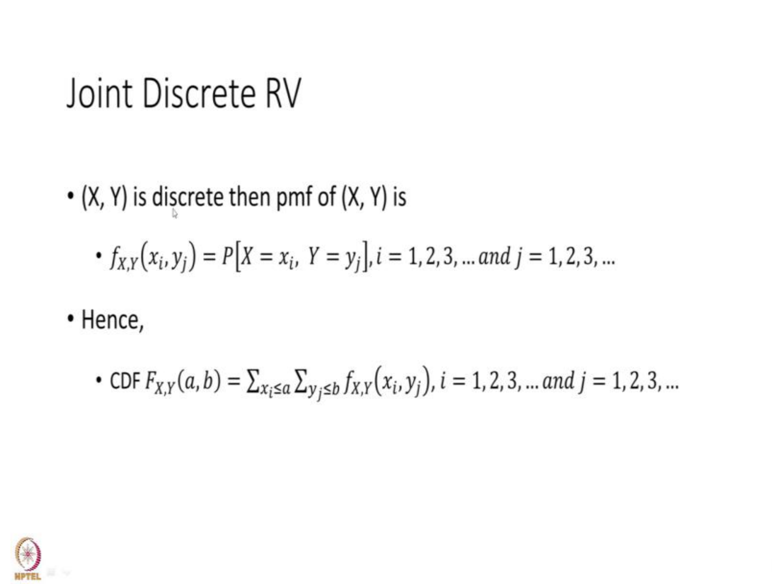
The marginal of CDF is defined as

$$F_X(x) = F_{X,Y}(x,\ \infty)$$

$$F_Y(y) = F_{X,Y}(\infty,\ y)$$

It means you take all possible values of Y and similarly you have a definition F of Y that is the marginal density of random variable Y as a joint distribution function of sorry I said marginal distribution function of Y is defined as a joint distribution function of XY where X takes the all possible values therefore it is shown here as an infinity. We will go into the detail definition of this in future.

(Refer Slide Time: 08:41)



So, if X and Y is a discrete joint random variable then the probability mass function of this XY is defined as

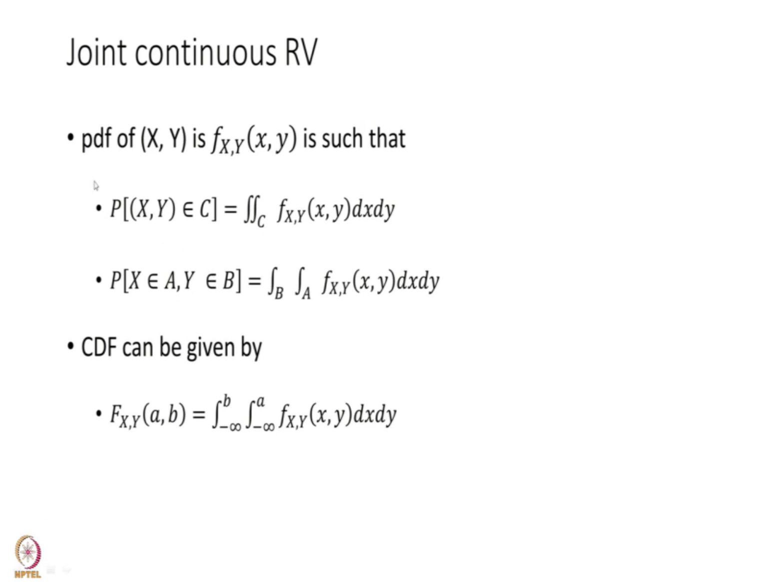
$$f_{X,Y}(x_i, y_j) = P[X = x_i,\ Y = y_j], i = 1, 2, 3, \dots \, and \, j = 1, 2, 3, \dots$$

Therefore the CDF the cumulative distribution function of the joint random variable discrete random variable XY at ab is defined as

$$F_{X,Y}(a, b) = \sum_{x_i \leq a} \sum_{y_j \leq b} f_{X,Y}(x_i, y_j),\ \ i = 1, 2, 3, \dots \, and \, j = 1, 2, 3, \dots$$

That is the discrete probability mass function of X joint random variable XY.

(Refer Slide Time: 09:39)



In case of continues now it is very obvious to you, pdf of a joint continues random variable XY is defined as small f of xy, is such that probability of XY belonging to a set C is

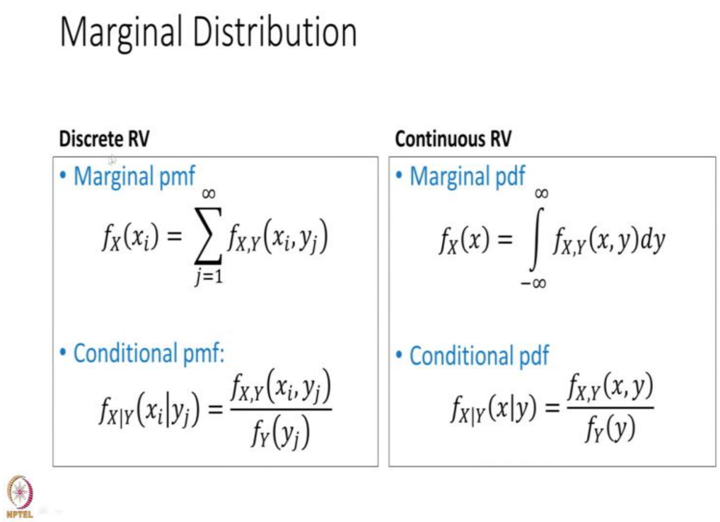
$$P[(X,\ Y) \in C] = \iint_C f_{X,Y}(x,\ y)dxdy$$

$$P[X \in A, Y \ \in B] = \int_B \int_A f_{X,Y}(x, y)dxdy$$

And then the CDF can easily be defined because in that case

$$F_{X,Y}(a,b) = \int_{-\infty}^{b}\int_{-\infty}^{a} f_{X,Y}(x,y)dxdy$$

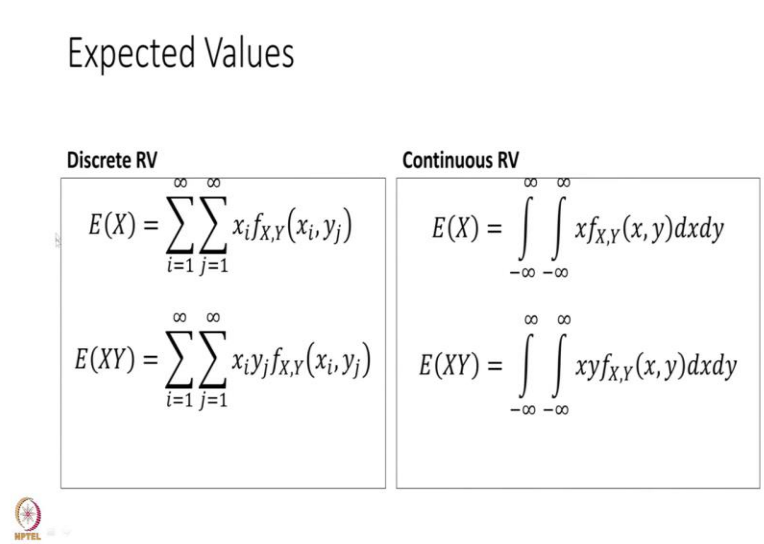
(Refer Slide Time: 10:51)



Let us define marginal distribution in a more with more clarity marginal distribution and also conditional distribution functions or conditional probability mass function. So, here both the cases are shown in case of discrete random variable marginal of any random variable x is defined by integrating out or summing out on all possible value of the other joint random variable. So, here the random variables are X and Y so if you are looking for a marginal for X you have to sum it up you have to in the case of continues you have to integrate over Y, here you have to sum it up over the all values of y_j.

Conditional pmf just as we define a conditional distribution function it is a joint distribution function or joint probability mass function divided by its marginal on which you are conditioning upon. So, it is a marginal of random variable y, in both the case you can make out as to if it is continues, how you define it and if it is discrete how you define it, it is very similar in nature.

(Refer Slide Time: 12:13)



The expected values are also defined accordingly in case of discrete random variable expected value of x is

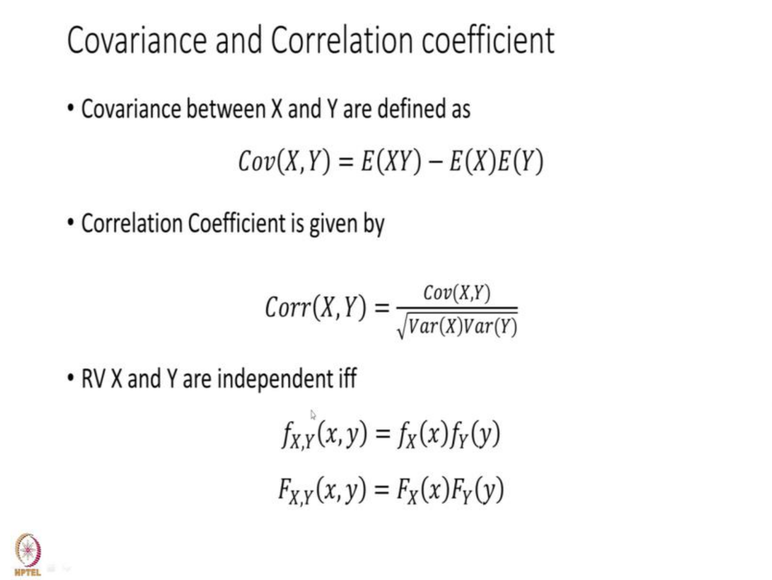
$$E(X) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i f_{X,Y}(x_i,\, y_j)$$

Otherwise you integrate it over the that is you fully integrate over Y and X but you multiply with the value of X the probability mass function.

$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,\, y) dxdy$$

Similarly if you have XY you have to put x_i y_j and here you have to put XY in the case of continues and continues random variable and then you have the expected value of X and Y and you can make it now what is the expected value of g of X or what is the expected value of another function H of X and Y both. They all can be derived from this particular method.

(Refer Slide Time: 13:18)



The, there are certain coefficients that we are interested in, one is the co-variance of X and Y, and the other is correlation coefficient between X and Y, please recall we have done the same thing in descriptive statistics. This we are doing with a general random variable, descriptive statistics has dealt with a data, now we are dealing not with any specific data but with a specific random variable which can take any value which is the your data in future you are going to call those as a sample values and these are going to be the actual random variable functions.

So co-variance between X and Y is defined as co-variance XY is equal to

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

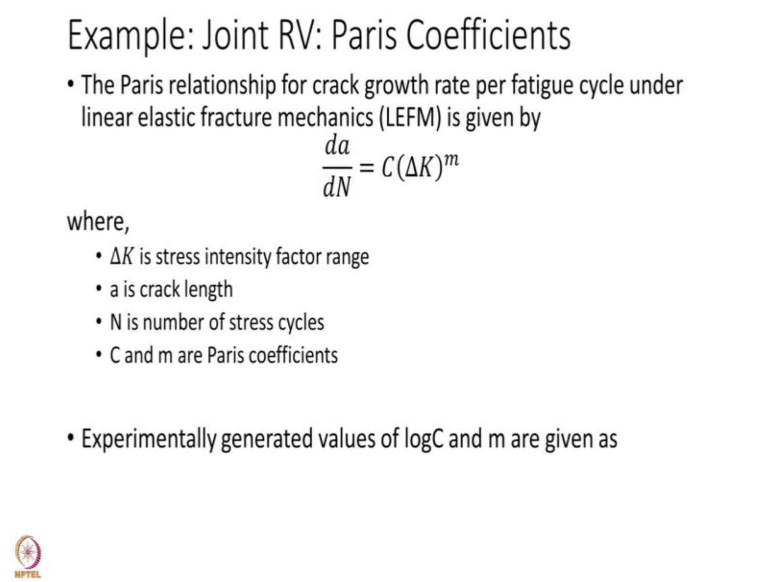
The correlation coefficient is defined as

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Now, we say that X and Y are two independent random variable if and only if the joint density function or probability mass function is actually a product of two separate marginal of X and Y, this is true both with respect to continues random variable and a discrete random variable.

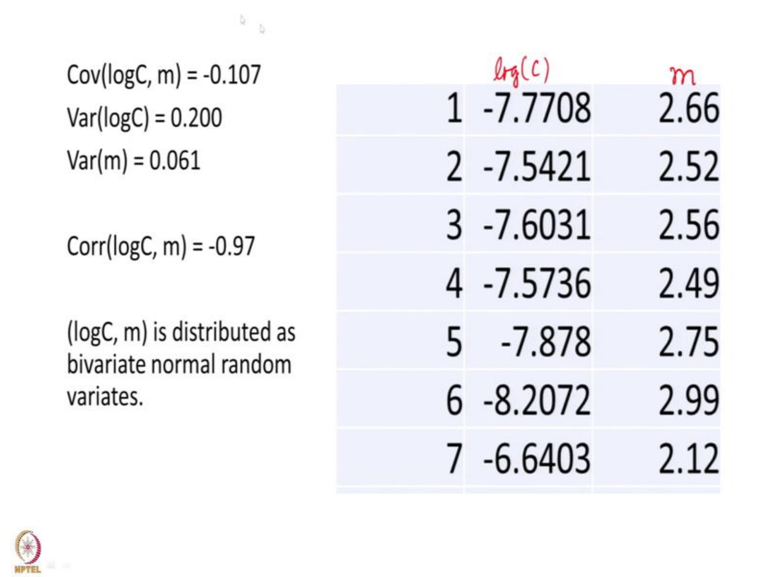
So in this slide I am not distinguishing between the two and in the same way the cumulative distribution function of xy joint XY is a multiplication of two cumulative distribution function of X and cumulative distribution of Y. If this happens then X and Y are called independent of each other. If X and Y are independent then also this happen that is why this condition of if with double f which means that is if and only if both of this implies each other.

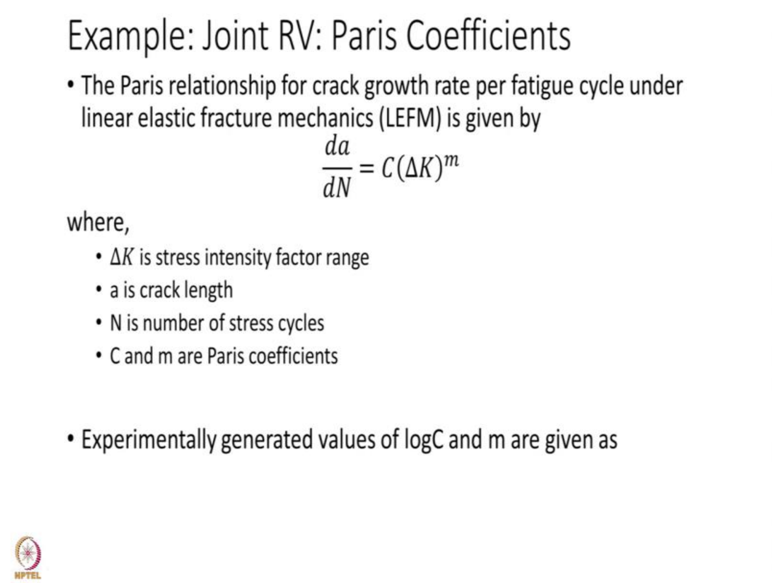
(Refer Slide Time: 15:40)



Let us take an example of joint random variable which we come across in the material science and metallurgy. Let us take the Paris relationship of crack growth rate per fatigue cycle under linear elastic fracture mechanics and that is given by this equation which is where a is a crack length, N is a number of stress cycle so it means that da by dN is the rate of growth of fatigue crack as a per fatigue cycle is equal to a constant Paris coefficient C stress intensity fracture range delta K to the power m.

(Refer Slide Time: 16:41)





Now we experimentally generated 7 such crack growth rate curves and we found the you will know in future that this can be solved by what is called a lock transformation through linear regression and you can find the values of logC and m that is log of the first coefficient of Paris coefficient and m. these values are found as covariance of log of here it is the data, let me first show you the data, these are the seven data points we have. I have not shown it properly let me write it down here so that it is clear to all of us.

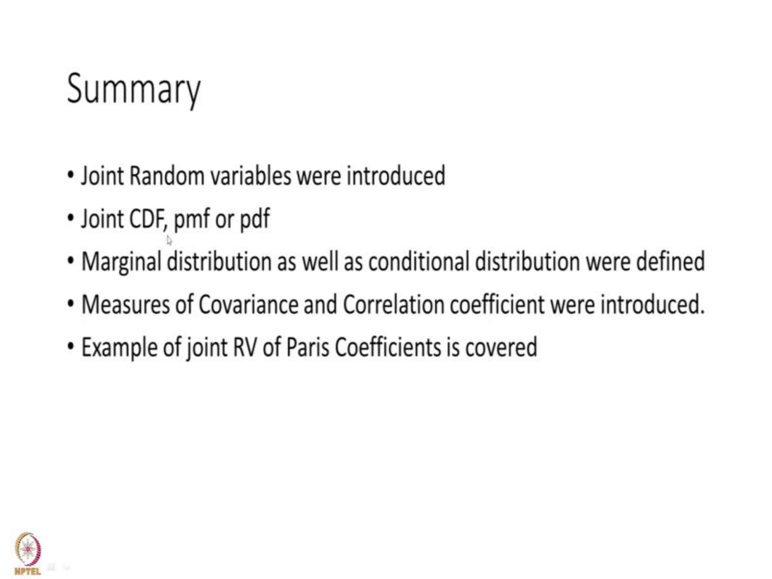
This is log of C and this is m, this information is missing and therefore you can see that here with this data if you try to find the co-variance of log C and m it is negative 0.107, the variance of logC turns out to be point 2, the variance of m turns out to be point 061. And the correlation coefficient turns out to be negative 0.97. It means that they are very closely correlated but negatively that is when m increases log C decreases this is the relationship and in fact this is known to be distributed as a bi-variate normal distribution.

We have not yet introduced the distribution functions and the special distribution as normal distribution but just for your information that bi-variate normal is this like a 2 variable normal distribution and this 2 random variables are logC and log m. Please not that this though Paris coefficients as they are called they are constants but we have to remember that when you estimate them for different fatigue crack growth curves that is for each of this you generate different crack growth curves they come out different because each estimate becomes a random variable.

From each variable you get some value of C and some value of m and they tend to vary and therefore they appear to be a random variable and here I am showing that they are a random variable in some sense. The estimated value please remember the, in the Paris equation it is not said that these C and m are random but when you actually perform the experiment there is random error into it which gets reflected as a these different values of logC and m and they become the random because each experiment is a random experiment and therefore these are random manifestation of logC and m and that is what I am saying they are highly correlated.

As we can expected because they are not suppose to be random but they are but from the random experiment we are getting different values so there is, so you know with when you perform an experiment there is always a little randomness in it and that gets reflected in this estimated value of coefficients and this is what I am showing, they are jointly distributed because this is not 0, the co-variance is not 0 and therefore the two random variables are not independent and therefore they are dependent random variables and they are a jointly distribution random variables.

(Refer Slide Time: 20:51)



Let us quickly summaries, we have introduced the joint random variables, then we introduced joint cumulative distribution function in case of a discrete joint random variable probability mass function and the in the case of continuous joint random variable. We define a probability distribution function. Marginal distribution we define as well as conditional distribution, we define the measures of co-variance and correlation coefficient in case of two joint random variables and we gave an example of joint random variable which have been obtained as Paris coefficients in several random experiments of generating Paris curves.

Thank you.