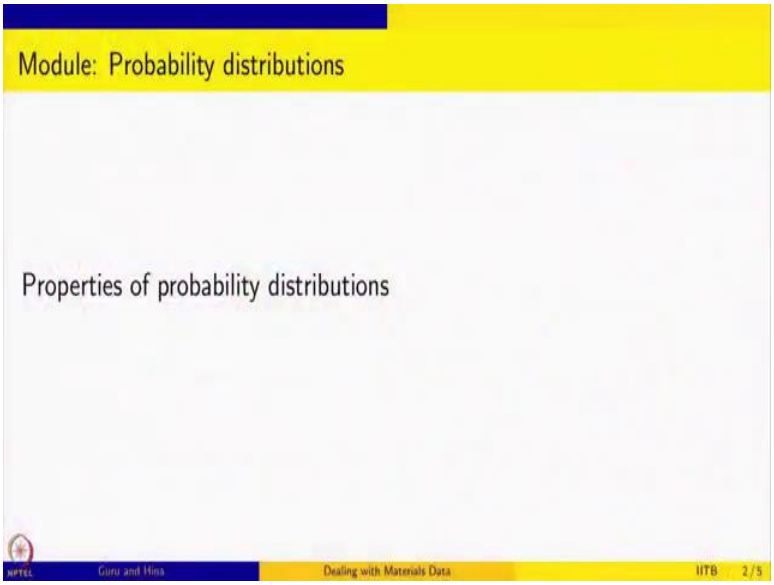
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor M P Gururanjan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 38
Properties of Probability distributions

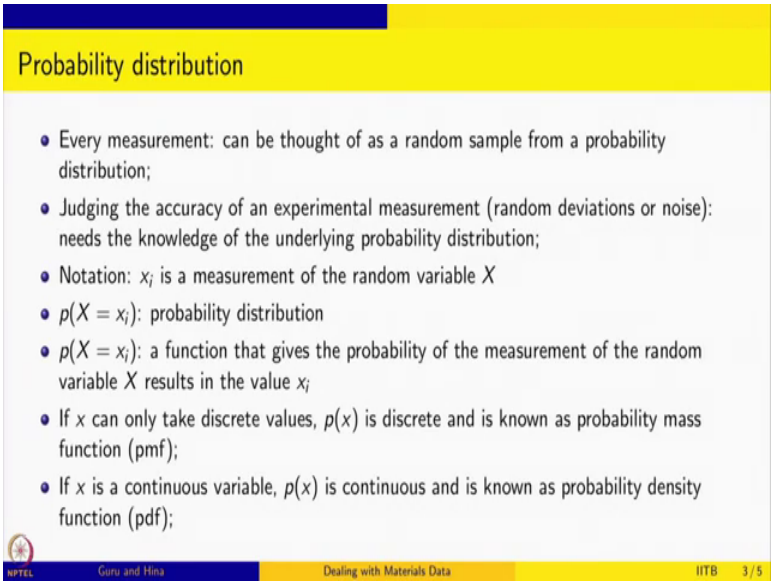
Welcome to the course on Dealing with Materials Data. In this course we are going to learn about collection, analysis and interpretation of data from materials science and engineering. We are in the third module. This is the module on probability distributions.

(Refer Slide Time: 0:30)



Specifically we are going to learn about properties of probability distributions in this session.

(Refer Slide Time: 0:37)



And one way to think about measurements that we do in the laboratory is that every measurement is a random sample from a probability distribution. For example, we have been looking at this case of conductivity of ETP copper, what we did is to make some 20 measurements and look at the data and found out how it is distributed, what its mean value is? What the standard deviations is? And things like that.

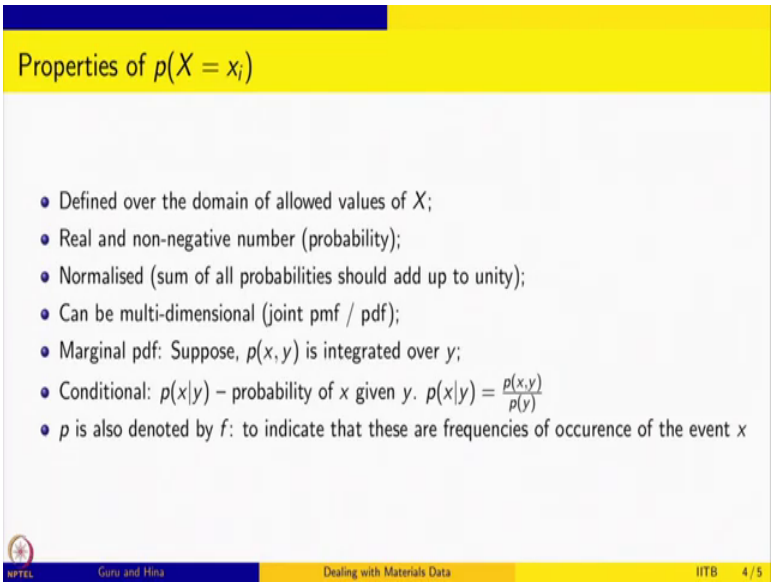
You can also think of the conductivity data to be normal distribution with a true mean and standard deviation and you can think of the different measurements we made as random sampling from this probability distribution.

So, this is another way of thinking about the experiment and in this scenario then judging the accuracy of the experimental measurement because we know that it should be a distribution like that and any variation from that because of random deviation is the noise and so if you understand the underlying probability distribution better, then we will be able to understand the accuracy better in our measurements.

We are going to use the following notation x_i, is a measurement of the random variable X. $P(X=x_i)$ is basically the probability distribution. So, it is a function that gives the probability of the measurement of the random variable X resulting in the value of small x in the i^{th} measurement so that is what this means. And if x can take only discrete values then p(x) is discrete and it is known

as probability mass function or pmf. And if x is a continuous variable, p(x) is continuous and is known as probability density function, pdf.

(Refer Slide Time: 2:31)



So, we are going to be dealing with both we will start with a discrete distributions and we will go to continuous distribution as we move along. Now, what are some of the properties of this probability distributions so they are defined over the domain of allowed values of X, outside of this domain typically they are assumed to be 0.

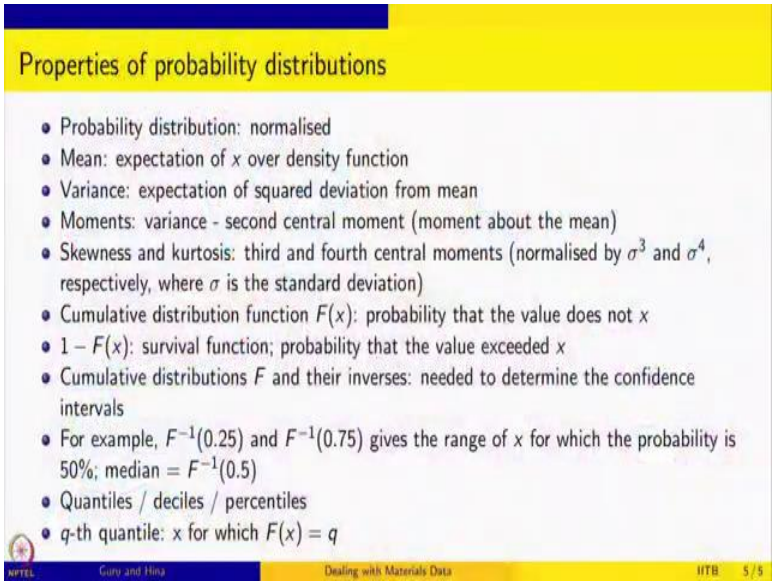
And this p(x) is a real and non-negative number, this is because we have already said that it is a probability so it has to be real, it has to be non-negative and because it is probability, because the probability of all the events should add up to one so it is also normalized so these values will lie between 0 and 1, so we will normalize it. And it can be multidimensional in which case you get a joint pmf or pdf so we are looking at one dimension x equals to x_i, it need not be you can have x, y, z, etc.

So, multidimensional are joint probability distributions are possible. When you have such multidimensional distributions sometimes you can define what is known as marginal pdf. Suppose, p(x,y) is a joint probability distribution for the variables x and y, if you sum or integrate over one of the quantities then you get the distribution function as a function of only one of the variables, this is known as the marginal pdf. So, it basically become the independent of the second variable.

You can also get what is known as a conditional probability distribution so p(x|y) so the pipe symbol basically is stands for given so given y what is the probability of x and that is given by there is a formula, so p(x|y) = $\frac{p(x,y)}{p(y)}$. You can also calculate p(y|x) and that will also be joint $\frac{p(x,y)}{p(x)}$. In all this we are assuming that p(x) and p(y) are not zero otherwise you cannot divided by p(x) or p(y), so that is also important.

And this also tells you that if p of x given y happens to be just p of x then x and y are independent, so you do not have to worry about the condition that y is given and in those cases you can also see that the joint probability distribution p of x, y will become p of x, p of y. Sometimes p is also denoted by f and we will also do it sometimes and that is to indicate that these quantities are sort of frequencies of occurrence of the event x, so we can also interpreted as the frequency of occurrence of any given event on the x.

(Refer Slide Time: 5:43)



So, let us continue to look at some of the properties of probability distributions. Like we mentioned earlier probability distributions are normalized, the mean value is the expectation of x over the density functions we have said that p of x is basically the probability of the random variate picking that value x, so if you take all those values and all those probability, multiply it by the value itself and sum or integrate then you get what is known as expectation and that happens to be the mean value.

And the variance is the expectation of squared deviation from mean, so you take the value, you take the difference of it with the mean and you squared it and you can take an expectation for this quantity, then you get what is known as variance. We have seen this in one of the previous sessions also where we talked about moments, so variance is basically a second central moment that is it is a moment about the mean and we have also looked at the skewness and kurtosis, these are third and fourth central moments.

And they are normalized by sigma cube and sigma power 4, where sigma is a standard deviation, so this we have looked at in one of the earlier sessions, we have defined this skewness and kurtosis and so they are defined for the probability distributions. We have also looked at cumulative distribution function so this is one of the things when we did, when we did the descriptive statistics.

So, from the empirical data we have looked at the cumulative distribution also you can also define cumulative distribution for any given probability distribution and that is denoted by capital F of x. Capital F of x basically gives the probability that the value is the, the cumulative probability, that the value is does not exceed x is what it stands for. And 1 minus F of x is known as the survival function, it is the probability that the value actually exceeds x.

So, F of x is cumulative distribution function 1 minus F of x is the survival function. Cumulative distribution functions and their inverses are needed to determine the confidence intervals. When we look at parameter estimation or hypothesis testing we will see that these are important. For example, F inverse of 0.25, what does that mean, F of x equals to 0.25 so it gives you the x value for which the probability does not exceed 0.25 and similarly F inverse of 0.75 gives it for the third quantile.

So, if you calculate this two values, so the inverse function basically tells you the x range for which the 50 percent of the probability is 50 percent or the data will fall within that 50 percent. Median, for example, is F inverse of 0.5 because F of x equals to 0.5 that is the probability cumulative probability of finding the value to be half is the F of 0.5. So, F inverse F of x equals to 0.5 so F inverse basically gives you that x value for which this happens which is the median.

So, similarly you can define quantile and deciles and percentiles and so on. For example, q-th quantile is that x for which F of x is equal to q so, F inverse of q is then x. So, these are some of

the properties of probability distributions and then we are going to look at each one of the distributions that we have mentioned some discrete like Binomial, Poisson and so on and some of them are continuous like normal Chi square, F, T and so on.

So, we are going to look at all this distributions and we are going to learn how to work with them. And how to generate some of these quantities? So, we are interested in generating the density functions, cumulative distribution functions and quantile functions which are basically the F inverses, the inverses of the cumulative functions as the less generating the random variates. So, that is what we want to do in the following sessions and we will start with the discrete probability distribution in the next session, thank you.