Dealing with Materials Data: Collection, Analysis and Interpretation

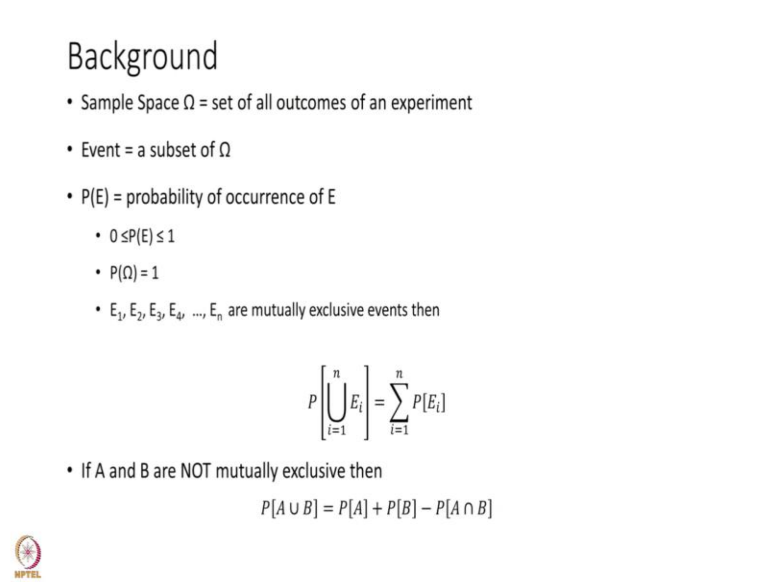
Professor. Hina. A Gokhale
Department of Metallurgical Engineering and Materials Science,
Indian Institute of Technology, Bombay.

Lecture 04

Random Variable and Expectation

Hello and welcome to the course on Dealing with Materials Data. In this session we will introduce what is known as random variable and its expectations in statistics.

(Refer Slide Time: 00:32)



First we will have a brief background there is something called as sample space when you when you perform and experiment it has several outcomes so set of all possible outcomes of an experiment is called as sample space.

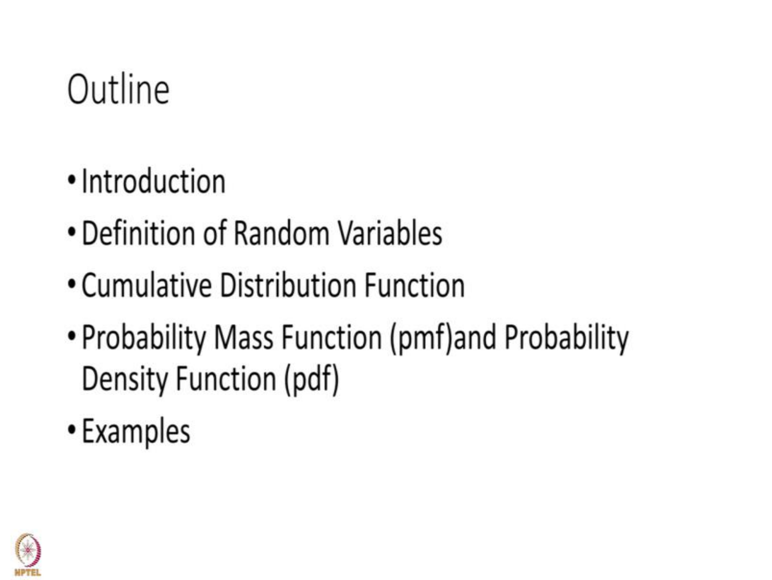
Any subset of this sample space is called an event in the terms of probability and probability of an occurrence of an event is define as probability of E and this probability has three qualities it is defined in with three qualities. One is that it is between zero and one. The probability of whole sample space is one and if you have mutually exclusive n events then the probability of unions of all these events is the summations of all these probabilities.

This also it can be derived from this that if A and B are not mutually exclusive events then the probability of then the probability of A union B can be written as

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

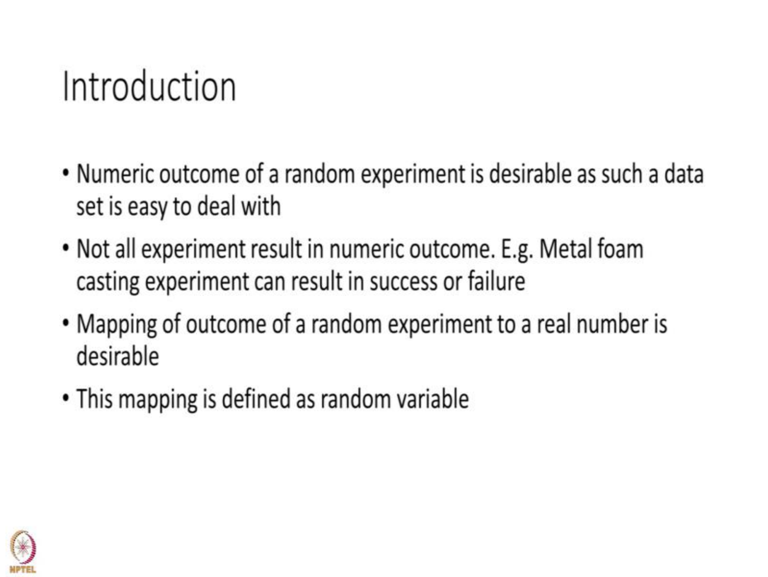
I think this background you already have.

(Refer Slide Time: 02:14)



Now let us move on so in this particular lecture we are… I am first going to make an introduction. Why do we need a definition of a random variable? Then we will define the random variable. Every random variable has one quantity attached to it which is called cumulative distribution function. Then we will define two kinds of random variables one is discrete and the other is continues. For discrete we will define the probability mass function and for continuous random variable we will define probability density functions and at the end we will give some examples.

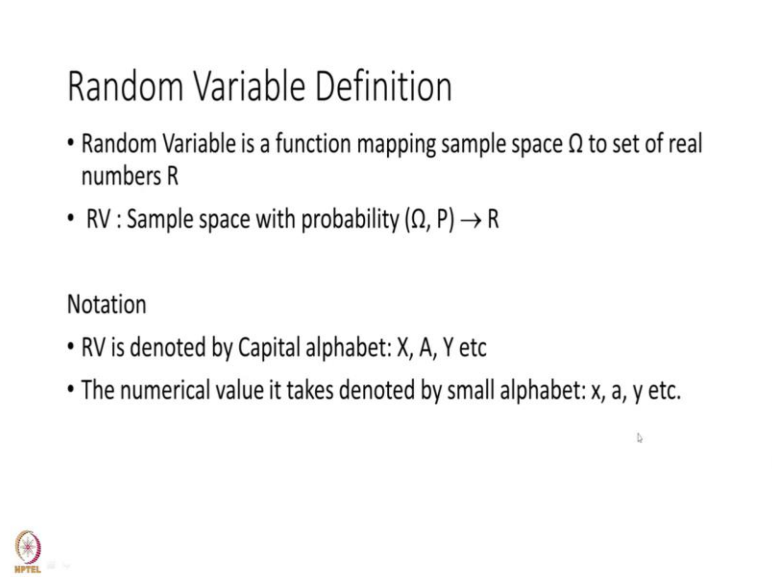
(Refer Slide Time: 02:56)



So let us move on. We perform a lot of experiments and the experiments may or may not have the result which is numerical. For example if you are performing a casting experiment to cast a metal foam sometimes the metal casting may form and make into a form a casting. And sometimes it just would not form and something would have gone wrong in the experimentation. So in that case the experimentation results in success or failure. So it is not necessary that every time the experiment results in a numeric data.

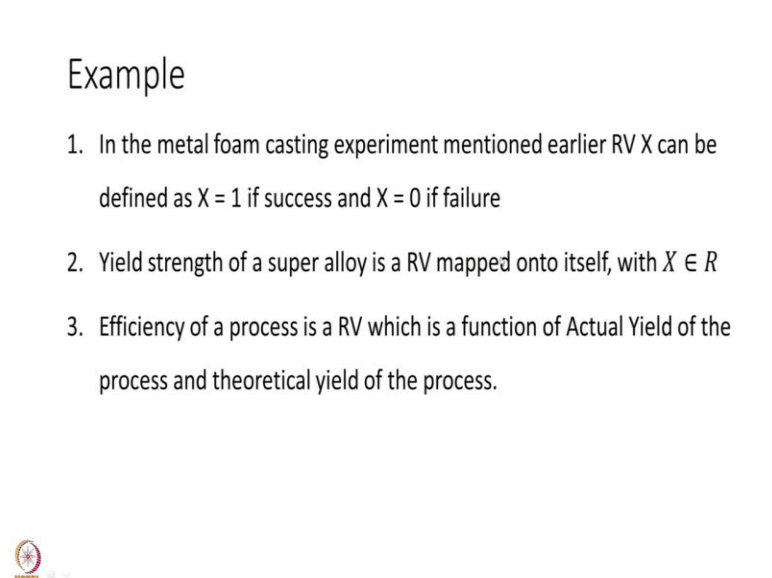
However having a numeric outcome or mapping the outcome into a numeric value is of immense help and it is desirable because that way we are able to do a lot of analysis on the data and therefore this mapping that we say that it is desirable. A mapping of desirable outcome of a random experiment to a real number is desirable is what we call a random variable.

(Refer Slide Time: 04:18)



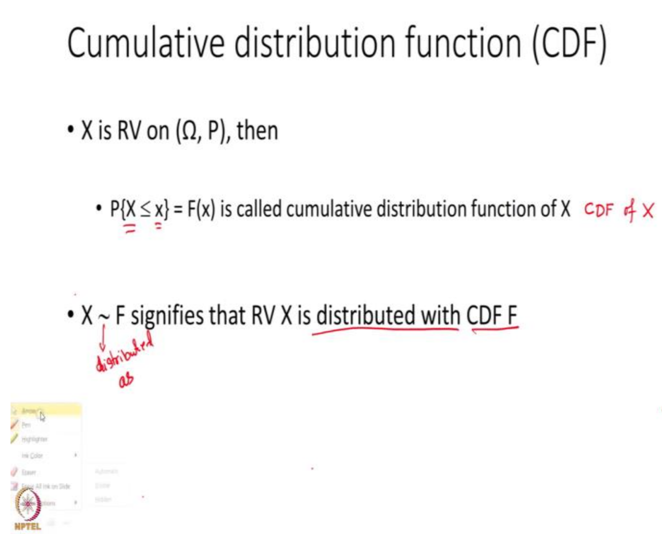
So let us formally define it. Random variable is a function mapping a sample space omega to a set of real numbers R. So RV is a random variable short form is a sample space with a probability P, map into a real line real number R. The notation now on we are going to uses all capital alphabets X, A, Y etc will be called a random variable and actual numerical values that it takes will be denoted by small alphabet such as given here.

(Refer Slide Time: 05:00)



So as I said in the case of metal foam if it is a successful experiment we call that X, random variable X is equal to 1 and if it is a failure we say that the random variable takes the value 1. A yield strength of a super alloy or yield strength of any alloy or any metal is a random variable mapped onto itself because yield itself numeric. Efficiency of a process is also a random variable which is a function of actual yield of the process and the theoretical yield of the process. So for example if you take a ratio of the actual yield of the process and the theoretical yield of the process and you call it efficiency that is also a random variable.

(Refer Slide Time: 05:54)



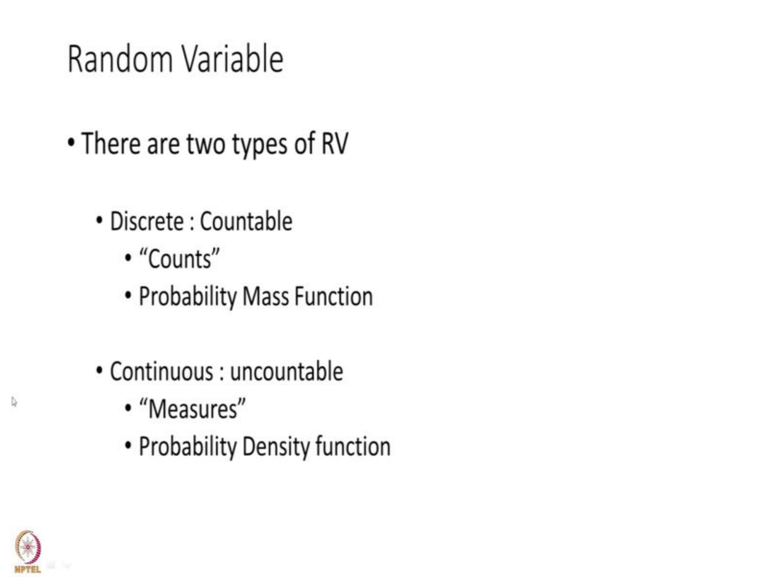
As I said before with all the random variable because it is based on a it is mapped from the sample space omega to a real line there is always a quantity which is called cumulative distribution function attached to any random variable. So if X is a random variable defined on a space probability space of X is RV on (Ω, P), then. Then probability that

- **$P\{X \leq x\} = F(x)$ is called cumulative distribution function of X**

F(x) is called cumulative distribution function of X or CDF in short, or CDF of random variable X.

Please note that this is a small letter and this is a capital letter so this is a value that X takes. So a notation for this is used is X is distributed as this signifies distributed X distributed as distributed with CDF F. So this is the notation. Let us move on.

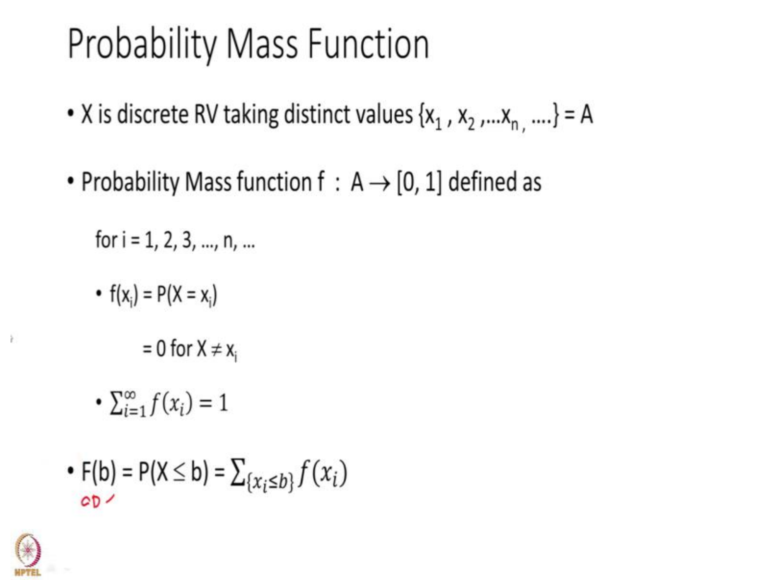
(Refer Slide Time: 07:32)



There are two types of random variables one is countable and the other is uncountable or continuous. What do we mean by that? Consider the experiment in which you are counting the number of defective outcomes that come out of a processing unit. So these are going to be numbers like there are two defectives three defectives five defectives etc. So these are called countable and with a countable random variable which is called discrete random variable there is a quantity attach to it which is called probability mass function.

While continuous is like yield strength is a continuous random variable and it actually refers to the measures. It measures the certain quantity and therefore it generally is a continuous random variable and with a continuous random variable probability density function is attached.

(Refer Slide Time: 08:40)



What is probability mass function?

- **X is discrete RV taking distinct values $\{x_1, x_2, \ldots\ldots\ldots x_n, \ldots\} = A$**

  Probability Mass function f: $A \to [0, 1]$ defined as

  For i = 1, 2, 3, …, n, …

  $f(x_i) = P(X = x_i)$

  $= 0$ for $X \neq x_i$

Well If X is a discrete random variable taking distinct value x1, x2, x3 ....xn ...so on and so for. Remember that it is a discrete so it takes a countable number of values. You call this all the values so it take as an A. Then probability mass function is a function defined a small f define from A to [0, 1] as f of xi is equal to probability that X takes a value of xi and it is 0 if x does not take a value of xi.

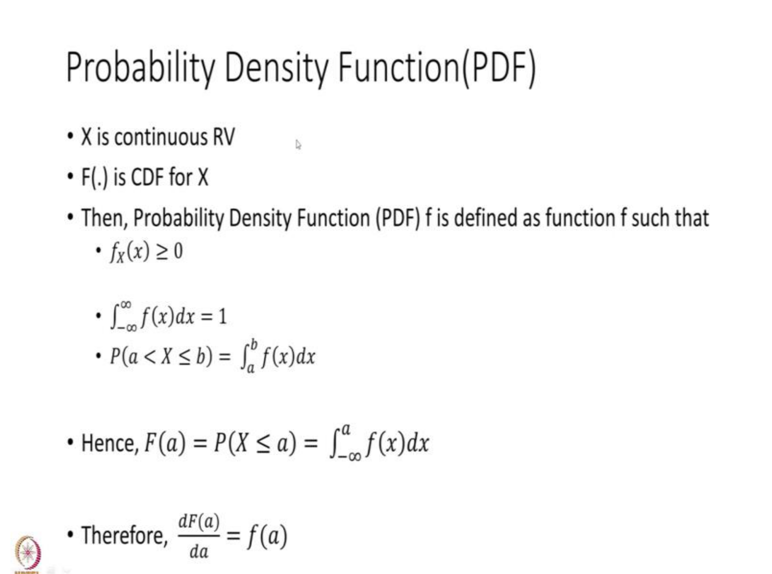
- $\sum_{i=1}^{\infty} f(x_i) = 1$
- $F(b) = P(X \leq b) = \sum_{\{x_i \leq b\}} f(x_i)$

Because that the complete set occurs is 1 and therefore it is the probability mass function says that summation of all the values of probability mass function that it can take is one. Summation if I have made a mistake I correct myself. Summation of all the values that a probability mass function can take is one and the cumulative this is please note that this is again the cumulative distribution function this is CDF. So this CDF of the random variable X is probability that

$$F(b) = P(X \leq b)$$

X is less than or equal to b in terms of probability mass function it can be written as summation over all the excise which are less than b, f(xi).

(Refer Slide Time: 10:34)



Probability density functions, suppose X is a continuous random variable and F capital F is a connected cumulative distribution function of X. Then the probability density function small f is define as a function such that it's always it takes on positive values its integration over all the possible values of X.

$$f_X(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a < X \leq b) = \int_a^b f(x)dx$$

Hence,

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

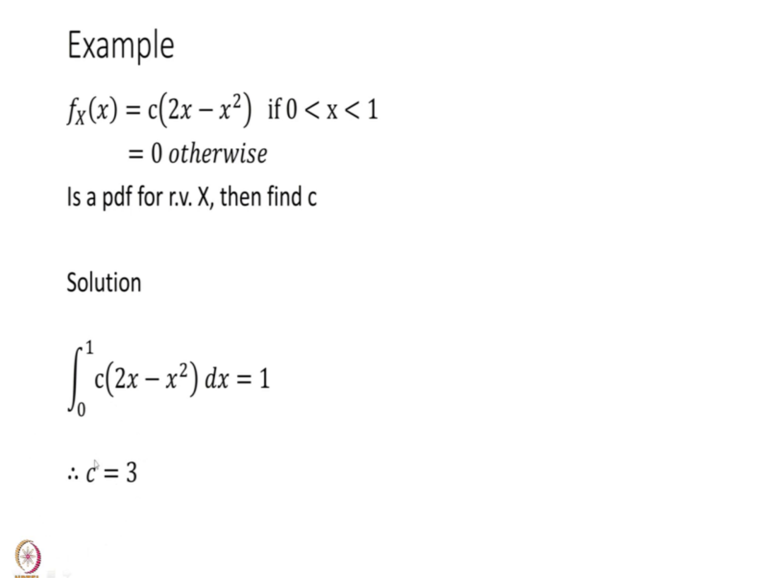
where f(x) is probability density function.

And from this you can say that the derivative of cumulative distribution function with respect to its argument a gives you the probability density function in case of continuous random variable.

Therefore

$$\frac{dF(a)}{da} = f(a)$$

(Refer Slide Time: 12:09)



Here is an example these are very simple examples suppose I say that

$$f_X(x) = \mathrm{c}(2x - x^2) \quad \text{if } 0 < \mathrm{x} < 1$$
$$= 0 \; otherwise$$

It's a pdf that is probability density function of x then find a c. The solution can be finding because the property that over the complete range of x which is here [0, 1], otherwise f(x) is 0.

$$\int_0^1 \mathrm{c}(2x - x^2)\, dx = 1$$
$$\therefore c = 3$$

The probability density function the integral is one and therefore you can easily show from simple algebra and arithmetic that this is actually three.

(Refer Side Time: 13:05)

Example

$$f_X(x) = c(2x - x^2) \quad \text{if } 0 < x < 1$$
$$= 0 \ otherwise$$

Is a pdf for r.v. X, then find c

Solution

$$\int_0^1 c(2x - x^2)\, dx = 1$$

$$\therefore c = 3$$

So with this we cover up what we summarize what we have covered today. We have introduced probability space which is a sample space with a probability measure. On the probability space we defined a random variable which is from probability space to a real line. With every random variable whether discrete or continuous there is cumulative distribution function attached which is f of x which is equal to probability of X less than or equal to x.

There are two types of random variables that we encounter in real life one is a discrete the other one is continuous. The discrete random variable is generally has a nature it is countable and it has attached to it probability mass function which actually is defined by small f of xi. xi is a realization of random variable x that small f of xi that probability that x takes on the probability of xi, i is number 1, 2, 3 .....n onwards. The continuous random variable actually takes on uncountable values and therefore it has an attached to it a probability density function.

Which is define as the probability function that a is less than the random variable X is less than or equal to b is integral from a to b f of x dx and of course f of x dx is a non-zero entity and if you integrate over the complete range of X it will be one that is it's another quality and if you take an integral from minus infinity to a value b then it gives you a CDF. The cumulative density, cumulative distribution function of the random variable of continuous random variable X and therefore the in the case of continuous random variable the first derivative of the cumulative distribution function with respect to its argument gives you the probability density function, Thank you