PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

Week - 02

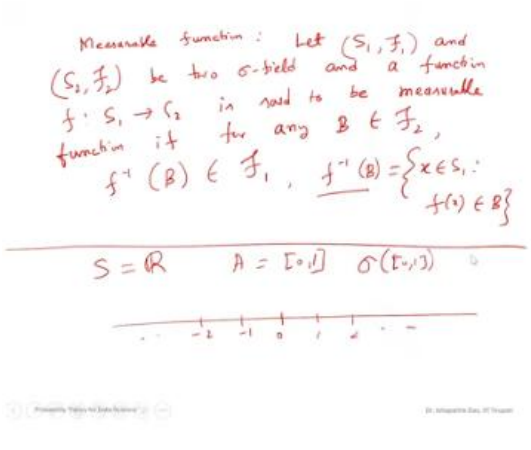
Lecture - 09

Events Defined by a Random Variable

So, let us consider the set of real numbers $\mathbb{R}$. So, the set of real numbers means you know that it contains natural numbers $\mathbb{N}$, whole numbers, integers $\mathbb{Z}$, both negative and positive integers, rational numbers $\mathbb{Q}$ (those can be represented by p/q), and also irrational numbers. So, combining all those, we form $\mathbb{R}$, the set of real numbers.

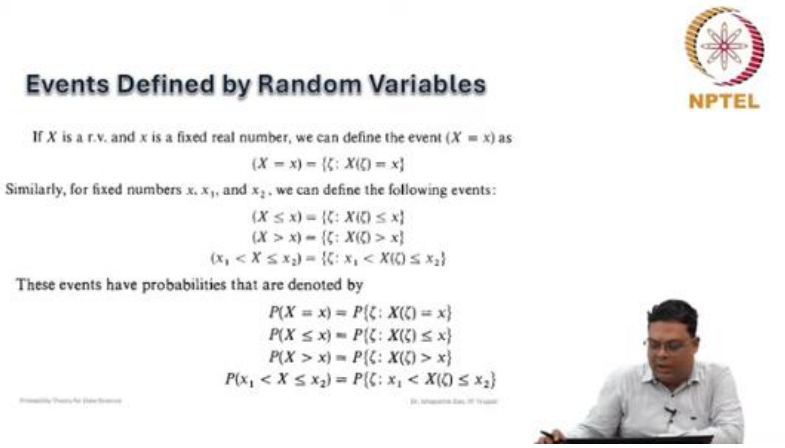
So, this real number we usually denote by $\mathbb{R}$.
So, this is 0, 1, 2, ... like this. This is -1, -2, ... like this. So, we usually denote it like this.

Now, in this set of real numbers $\mathbb{R}$, we want to take some subset. So, you can consider, suppose A = {0, 1}, this subset, and the σ-field generated by A. Suppose this is {0, 1}. So what is the σ-field generated by this? We have already discussed it. So, if you consider the whole set of real numbers $\mathbb{R}$, we need to satisfy these three properties. First, $\mathbb{R}$ has to be in the class of subsets. If any element is in this subset, then its complement also has to be part of this set.



Additionally, if you have any infinite collection in C, then the union of this countable infinite collection of A_i also has to be in C. So, if you consider S = $\mathbb{R}$, it is infinite because all the examples we discussed up to this point are finite sets. That is why you can

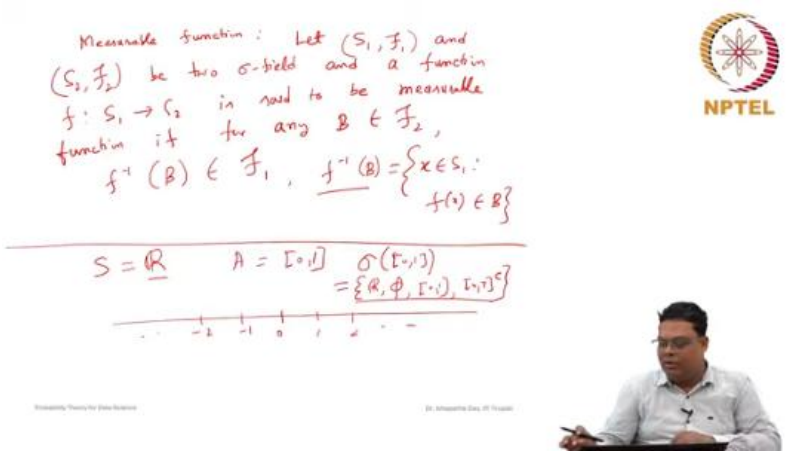
say the power set contains 2^n number of elements. But if it is infinite, like the real numbers $\mathbb{R}$ or even the interval [0, 1], then it is countably infinite. If you consider the power set, it is huge; you cannot just count it.



You can't say it's 2^n type of things. It is not finite, so we represent it with some cardinality. This is a different concept, and we don't want to go into that complexity. That's why we want a particular σ-field that is actually useful for us. Here, if you consider the σ-field generated by this set $\mathbb{R}$, we want to consider all subsets. The σ-field generated by the element 0 must include this set $\mathbb{R}$, and its complement, which is the null set $\emptyset$, must also be included.

Since we are saying the σ-field is generated by this set A, this set has to be included. Its complement also has to be there, which is the complement of the interval (0, 1). If you take the union, then this will be $\mathbb{R}$. So, this is the smallest σ-field generated by this set A. Now, if you consider some special subsets that are there in the real line...

So, let us consider the σ-field generated by two real numbers, A and B, such that $A \le B$. Now, consider a subcollection denoted as C, which consists of all A such that $A \in \mathbb{R}$ and $A < B$. If you take the σ-field generated by this collection, let's denote it as C_1. Now, we can take another collection, C_2, which consists of intervals from $-\infty$ to x, where $x \in \mathbb{R}$. Let us consider the σ-field generated by C_2.



We won't go into more details about it, but it can be proved that whether you take open or closed intervals, the σ-field generated by C_1 and the σ-field generated by C_2 will be the same. This is because all sets in C_1 will also be included in the σ-field generated by C_2. You need to include the complements and countable unions as well. So, both σ-fields are the same, and this type of σ-field has a known name. The σ-field generated by C_2 is known as the Borel σ-field.

The Borel σ-field contains all intervals, including half-open and half-closed intervals, as well as intervals from $-\infty$ to x. We denote this as $\mathcal{V}$, representing the Borel σ-field. This concept is very useful, and we will see how we utilize the Borel σ-field. Most cases will involve sets that are inside the Borel σ-field, although there are some cases that are not Borel measurable. We won't go into detail about that right now; we just want to understand the basics.

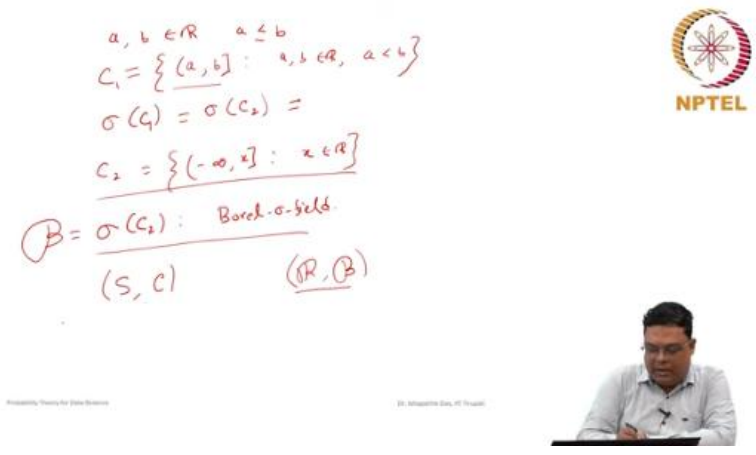
Now, if you consider a sample space, we are going back to the concept of a sample space in random experiments. This sample space may be any abstract set, not necessarily a set of real numbers or a subset of real numbers. You will have a collection of subsets of S that form a σ-field. This collection constitutes your probability space whenever we define a probability on it. In the real number system, we have real numbers and the Borel σ-field.

We want to relate these two things: on the left side, we have our usual sample space and some collection of subsets of S that form a σ-field; on the right side, we have the real number system and the Borel σ-field. We want to relate these through a measurable function. So, a measurable function relates X, which is from S to the set of real numbers.

Until now, whatever we have defined as a measurable function should be clear, but feel free to go back and review it. So, what is a measurable function? So, there are two σ-fields: $(S_1, \mathcal{V}_1)$ and $(S_2, \mathcal{V}_2)$. S_1 is a non-empty set, and $\mathcal{V}_1$ is a collection of subsets of S_1 satisfying some properties known as a σ-field. S_2 is another non-empty set, and $\mathcal{V}_2$ is a collection of subsets of S_2 that form a σ-field. A function f: $S_1 \rightarrow S_2$ is said to be a measurable function if, for any $B \in \mathcal{V}_2$, the inverse image $f^{-1}(B) \in \mathcal{V}_1$. Now, considering the Borel σ-field, it contains at least the sets generated by these kinds of sets.

You can see that by taking intersections, complements, and unions, all these things will be there. Any intervals will be included, and any singleton sets are also there. You can check that if you take the intersection of a closed interval, its complement is included, and if you take unions, those are included too. In that way, you can show that singleton sets like {1}, {2}, {3}, and {4} are also inside this Borel σ-field. Now, if you consider a measurable function X from S to $\mathbb{R}$, let me pause for a minute.

So, we want to define a random variable. A random variable is a measurable function. Let us recall again that S is the sample space and $\mathcal{C}$ is a collection of subsets of S that forms a σ-field. $\mathbb{R}$ is the set of real numbers, and $\mathcal{W}$ is the Borel σ-field, which contains the Borel sets. A random variable, denoted as X, is a measurable function from S to $\mathbb{R}$.



So, basically, X is a measurable function from S to $\mathbb{R}$, and we say it is Borel measurable. What does Borel measurable mean? For any set $B \in \mathcal{W}$ (Borel set), $f^{-1}(B)$ must belong to

$\mathcal{C}$. Here, $\mathcal{C}$ is the σ-field associated with the sample space S, and B is a Borel set. Thus, if $f^{-1}(B) \in \mathcal{C}$, then it is known as a measurable set.

Therefore, if X is a measurable function, it is called a random variable. Now, let us discuss some examples of random variables. So, consider a simple random experiment, such as tossing a coin. The sample space, S, consists of H (heads) and T (tails). We want to take the σ-field, $\mathcal{C}$, to be the power set.

This means $\mathcal{C}$ includes the empty set, S, H, and T, containing all subsets of S. Now, we can define a function. Let X: S → $\mathbb{R}$, defined as follows:

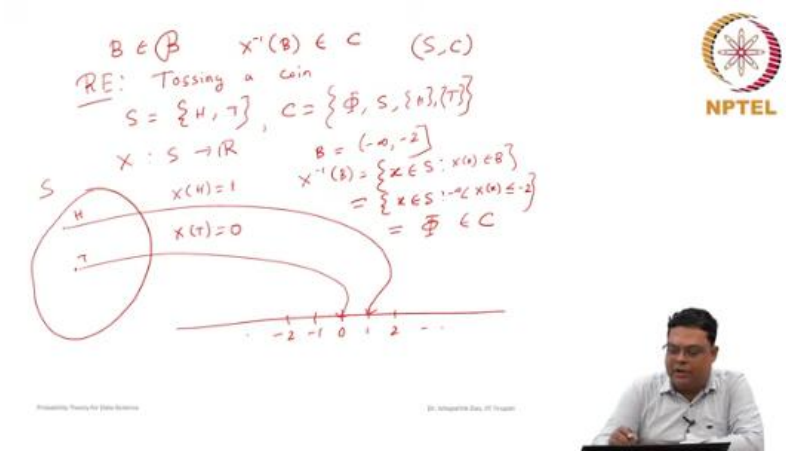
- X(H) = 1
- X(T) = 0.

We want to determine whether X is a random variable. First of all, it is a well-defined function.
To show that X is a random variable, it is enough to consider any Borel set. A Borel set is generated by all such sets. It suffices to show that the inverse of this set is in the σ-field. Let us consider one Borel set, B, which could be an interval. For simplicity, let's take the interval (-∞, -2).

So, what will be $f^{-1}(B)$? $f^{-1}(B)$ consists of all elements in S such that $X(x) \in B$. According to the definition of $f^{-1}(B)$, we check if X(H) or X(T) belongs to B. Since B is the interval (-∞, -2), we need to determine whether $X \leq -2$. The values of X are:

- X(H) = 1
- X(T) = 0.

Neither of these values satisfies the condition of being ≤ -2. Therefore, $f^{-1}(B) = \emptyset$. The null set $\emptyset$ is included in $\mathcal{C}$, confirming that this is true for any other value as well. Let us consider another set, B_2, which is defined as the interval (-∞, 0.5). This can be any real number, whether rational or irrational.



Now, what is $X^{-1}(B_2)$? $X^{-1}(B_2)$ consists of all $X \in S$ such that $X(x) \in B_2$. This means we are looking for all $X \in S$ such that $x > -\infty$ and $x \leq 0.5$. Next, let's check our mapping. We have X(head) = 1 and X(tail) = 0.

Since X(head) = 1, it does not satisfy the condition for B_2. However, X(tail) = 0, which does satisfy this condition. Therefore, $X^{-1}(B_2)$ includes tail, confirming that it is a subset of $\mathcal{C}$ and belongs to $\mathcal{C}$. Furthermore, we can consider other Borel sets. For instance, let's look at a singleton set denoted as {X = 1}.

This set is defined as all $X \in S$ such that X(x) = 1. In this case, since X(head) = 1, this set corresponds only to head. Thus, this singleton set is also included in the σ-field, $\mathcal{C}$. Similarly, if we consider the notation $X \leq 1$, it is denoted as all $X \in S$ such that $X(x) \leq 1$. Since there is nothing on the left-hand side, this indicates it is greater than $-\infty$.

This condition includes all values satisfying it: X(head) = 1 and X(tail) = 0, confirming that both are in the sample space. In essence, the collection of sets we are considering can be denoted as $\mathcal{C}_2$, which includes all intervals from $-\infty$ to x, where $x \in \mathbb{R}$. The Borel set, generated by $\mathcal{C}_2$, contains all such sets. If you take the inverse of any set, it will have the same meaning as the notation $X \leq x$, where $X \in S$. To avoid confusion, we need to clarify the notation we are using. Since we denote X as a real number, it's important to use distinct notation for elements in S.

Therefore, we should adjust our expressions for clarity. We denote S here, with the small

x representing a real number. Now, if you consider this type of set—$\mathcal{C}_2$, defined as the interval $(-\infty, x)$, where $x \in \mathbb{R}$—then B is the Borel σ-field generated by $\mathcal{C}_2$, which is just the Borel σ-field. So, what is $X^{-1}((-\infty, x))$? This is defined as all $s \in S$ such that $X(s) \leq x$, where $-\infty < X(s) \leq x$.

So again, if we use the notation $X \leq x$, this is just a simplified notation indicating that all $s \in S$ satisfy $-\infty < X(s) \leq x$. Both notations mean the same thing. Essentially, $X^{-1}((-\infty, x))$ is just what we denote by $X \leq x$. So, what we understood now is... So, let us recall from the beginning we started with a random variable.