PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

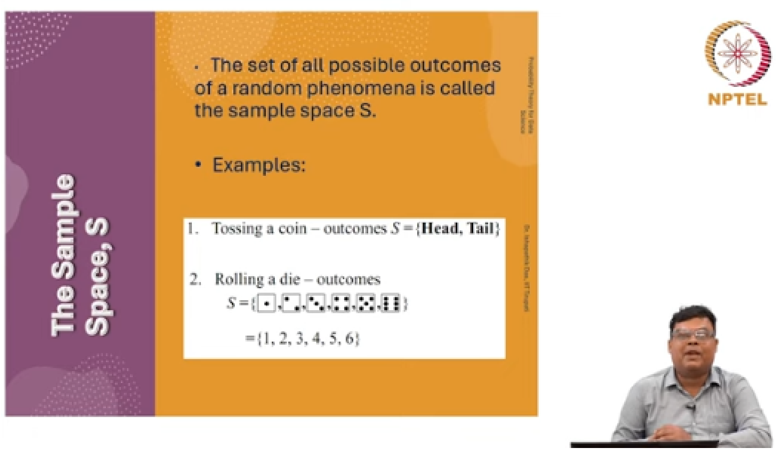
Indian Institute of Technology Tirupati

Week - 01

Lecture - 02

Sample Space and Events

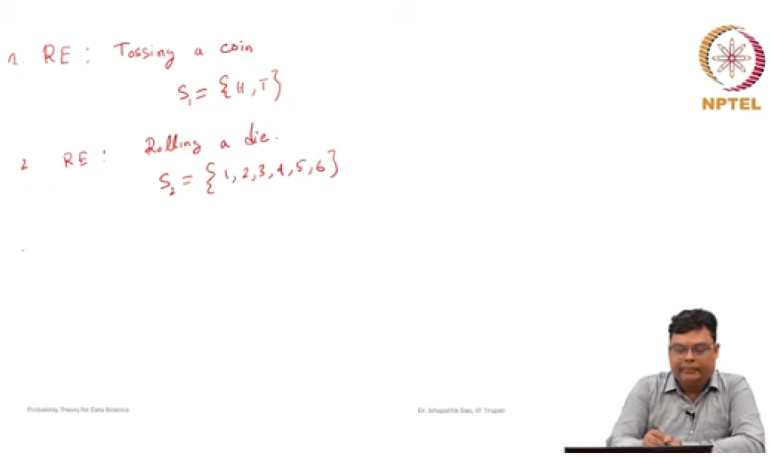
Let us discuss this sample space. What is sample space? Whenever we are doing a random experiment, meaning we are repeating this phenomenon in our lab, we know all the outcomes of this phenomenon. That random experiment, whenever we are doing, all the possible outcomes of a phenomenon or random experiment, we say all possible outcomes of a random phenomenon or random experiment are called the sample space. For example, if you are tossing a coin, suppose we denote that random experiment by this abbreviation RE.



Suppose the random experiment, it is tossing a coin. So, what are the possible outcomes of this random experiment? It may be head or tail. These are the possible outcomes. All possible outcomes of this random experiment, it is known as sample space.

Let us consider another random experiment. Suppose, this is random experiment second example. Suppose, throwing a die. Or rolling a die. Suppose you are rolling a die. Then, what are the possible outcomes?

S = all possible outcomes. Suppose you denote the sample space as $\mathbf{S}_1$ here, and the sample space $\mathbf{S}_2$. This is the set of all possible outcomes: $\{1, 2, 3, 4, 5, 6\}$. It seems like sample space contains always finite number of elements. This is like an abstract.



These are not real numbers, head and tail. These are some integer values. It may take any real numbers also. It may not be discrete. For example, let us consider the random experiment of the lifetime of an electronic group.

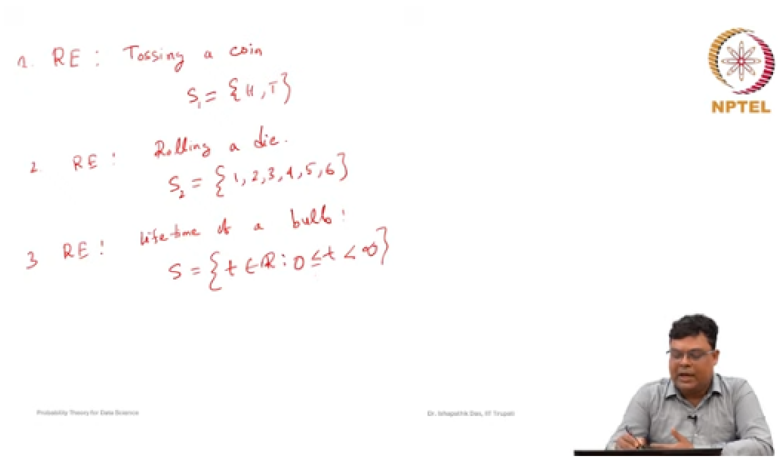
Suppose it is a transistor. So, the lifetime of electronic goods, or you can consider a very simple bulb, suppose. The lifetime of electronic goods. So, lifetimes mean we are measuring time. Suppose we consider one bulb and want to see how many years or how many hours, in some units.

It is actually, if you consider properly, time is a continuous variable. It may be any point on the real line, but if you use units like hours or minutes, then it will be discretized. However, if you consider it theoretically, it could last any time on the real line. So, we will see how much time **t** it is actually working. So, after that, it fails.

So, this sample space, it can be represented by a real number **t**. For example, if you buy the bulb and then it is not working immediately, then the lifetime will be zero. So, it is possible that theoretically it may be zero, otherwise it is greater than zero. And we don't know how long it will continue. We don't know if there is any maximum limit; it can work for more than **m**, or it is not possible. We do not know that much value.

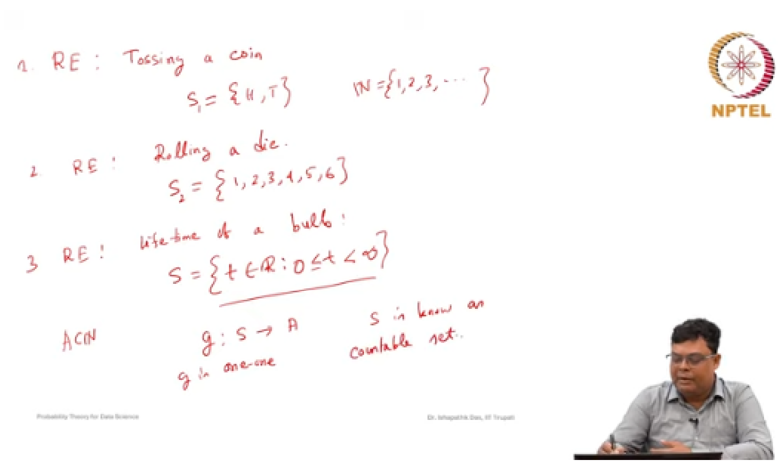
So, that is why theoretically we write that **t**, the lifetime of a bulb, is such that **t** belongs to the real numbers and **t** is greater than or equal to zero and less than infinity. So, here the sample space is uncountable. So, we know that countable set and uncountable set.

So, a countable set means any set where, if you have a subset of natural numbers, **A** is a subset of natural numbers. Natural numbers are denoted by 1, 2, 3, and so on.



A is a subset of natural numbers. If you have a one-to-one correspondence, then **A** may be the whole set of natural numbers also. If you have a bijective correspondence from **A** to some set **S**, then **S** is known as, so **g** is one-to-one, or we say that it is one-to-one. If there is any one-to-one correspondence from **A** to **S**, it has to be one-to-one also. In other words, if there is a one-to-one correspondence from **S** to **A**, where **A** is a subset of natural numbers, either **A** may be the whole set of natural numbers or a subset of natural numbers. If there is a one-to-one correspondence from **S** to **A**, or if there is a bijective function from **S** to **A**, then **S** is known as a countable set.

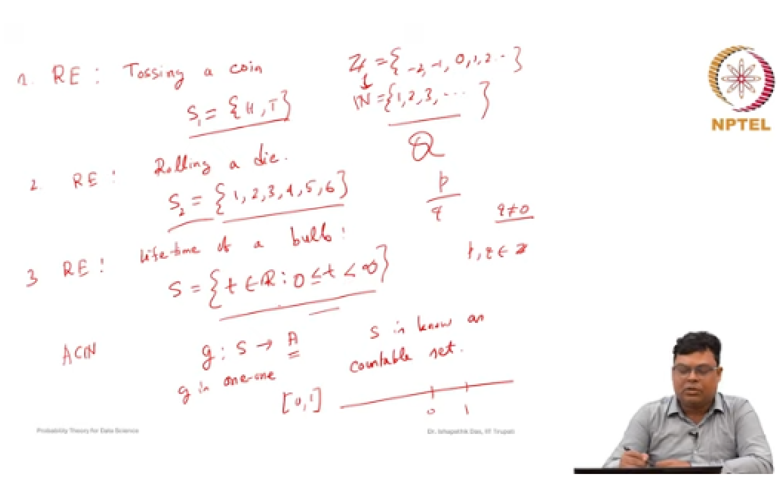
It may be infinite. So, for example, the set of natural numbers itself is countable because 1 goes to 1. It is a bijective function and one-to-one. And if you consider **Q**, **Q** is the set of rational numbers. That means all the elements can be represented by **p/q**, where **q** is not equal to zero and **p** and **q** are integers. Note that **Z**, the set of integers, contains 0, 1, 2, -1, -2, and so on.



There is also a one-to-one correspondence between **Z** and the natural numbers. Similarly, **Q** also has a one-to-one correspondence with the set of natural numbers. If it is a finite number, it is also a countable set because **A** is a subset of the natural numbers. So you can take this subset and then show that it is just an identity map. It is a one-to-one correspondence from the set to the subset of the natural numbers.

So this is also a countable set. These are all countable sets, but whenever we cannot represent a one-to-one correspondence between the set and a subset of natural numbers, then it is an uncountable set. For example, the set of real numbers. The real numbers are nothing but the union of rational and all the irrational numbers. So, now even if you consider the interval from 0 to 1 on the real line, all the points in between 0 and 1 form an uncountable set.

So, 0 to infinity is also an uncountable set. So, sample space may be a countable set, sample space may be an uncountable set also, it may be finite, it may be countably infinite, or it may be uncountably infinite. So, this is the sample space. Next, we will discuss an important topic known as an event. So, what is an event?



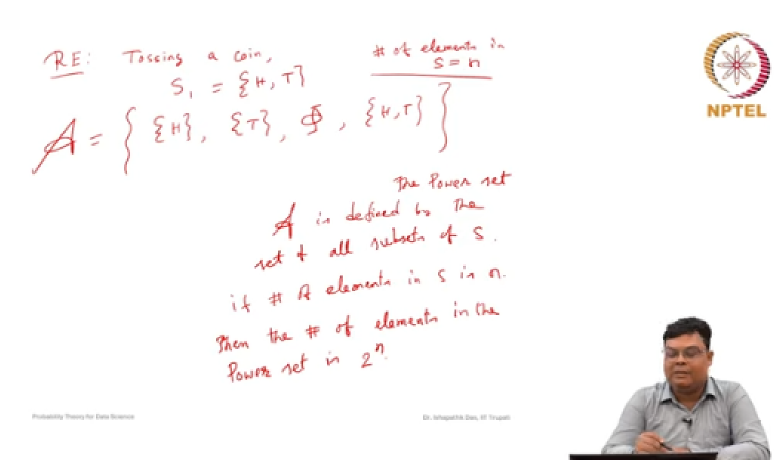
That is because later on we will discuss the probability of an event. So, that's why it should be understood very clearly. Many times, we just say, for a layman, it is nothing but a subset of a sample space. So, if **S** is discrete, you can consider all subsets of the set; these subsets are the events we discuss. So here, any subset of the sample space **S** is called an event. Even when **S** is a discrete set, the Venn diagram represents the event.

For example, if you consider the random experiment of tossing a coin, then in this experiment, the sample space $\mathbf{S}_1$ contains heads and tails. So, what are the subsets we can consider? Subsets include one subset like heads, another subset like tails, the null set, and the set **S** itself. So, containing all these subsets, all subsets of **S** are considered. Note that if it is a finite set, then we can count all subsets.

If the number of elements in S is, suppose n, then we say it is called a power set. So, we denote it like a script A. The power set is nothing but all subsets. So, the power set, we say, is defined by the set of all subsets of S. If you consider the set of all subsets of S, it is known as the power set.

If the number of elements in S is n, then the number of elements in the power set is 2 raised to the power of n. Note that this notation is just for a short type of writing. For example, if S1 contains only two elements, then the power set contains all the subsets of S1. It contains 1, 2, 3, and 4 subsets, which is 2 raised to the power of 2, or 4 elements. So, this shows the relationship between the power set and the finite set.

If it is infinite, then 2 raised to the power of infinity is not a number. If it is infinite, then 2 raised to the power of infinity is not a number. But if S is uncountable, then we cannot say that the power set is countable. If S is an infinite set, then the power set will be much larger. We represent this with a cardinal number.



For example, the cardinal number for natural numbers is denoted by eta. All sets with the same cardinal number as natural numbers are countable. If there is no one-to-one correspondence between a set and natural numbers, then it is uncountable. For example, the cardinal number for the set of real numbers is represented as 2 to the power of eta. This helps us understand larger sets because, with infinite sets, we cannot compare them just by the number of elements as we do with finite sets.

This set is bigger than this set because it contains four elements, while this one contains only two elements. We can compare these finite sets. But for infinite sets, like the set of natural numbers and the set of rational numbers, we cannot compare them just by counting elements. Both are infinite, so we cannot compare their sizes in the same way as finite sets. So, by using some bijective correspondence, or the concept of one-to-one mapping, we can have a way to compare infinite sets.

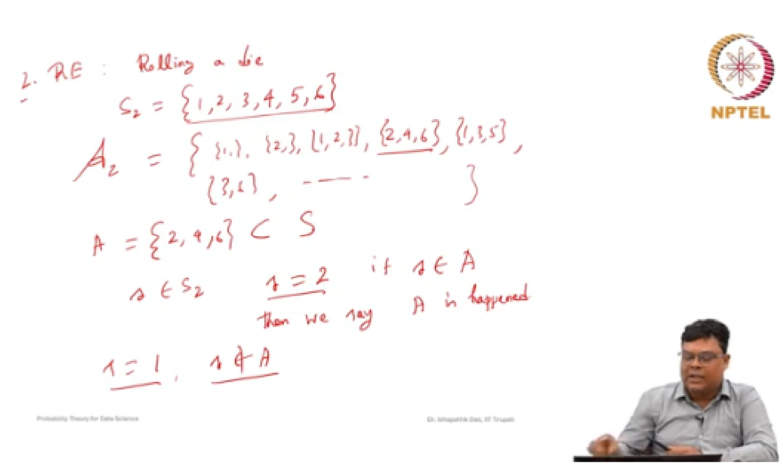
For example, eta is a cardinal number for natural numbers. All countable sets should have the same cardinal number because there is a one-to-one correspondence with natural numbers. And when we go to uncountable sets, like the set of real numbers, it has a cardinal number of 2 to the power of eta. These are some concepts that are not that much needed here. So, we will discuss the related things.

An event is a subset of a sample space. So, if A is a subset of the sample space, then A is an event. Similarly, if B is also a subset of the sample space, then B is an event. So, let us discuss another example. For example, let's consider a random experiment of rolling a die.

The sample space contains 1, 2, 3, 4, 5, and 6. So, the power set of S2 will contain many subsets, 2 to the power of 6 in total. Some of the subsets are: {1}, {2}, {1, 2, 3}, {2, 4, 6}, {1, 3, 5}, {3, 6}, and so on. Suppose we consider the subset A = {2, 4, 6}. This is a subset of S and belongs to the power set. A includes all the even numbers.

When we say this event occurs, it means that in one observation of rolling a die, the outcome will be one of the elements in A. For example, if the observed result is 2, then 2 is an even number and is part of the subset A. Now, if the outcome of the observation belongs to the subset A that we are considering, then we say that A has occurred, or A has been observed. This means the event A happened. So, whenever we see that the outcome is 2, and 2 belongs to A, then A has been observed.

Similarly, if the outcome is 4, A is observed. If the outcome is 6, A is also observed. However, if the outcome is 1 and 1 does not belong to A, then A has not occurred. In this case, we say that the event A did not occur. So, that is the concept of how we define an event. Now, if you consider the third random experiment, it is measuring the lifetime of an electronic good, like a bulb.

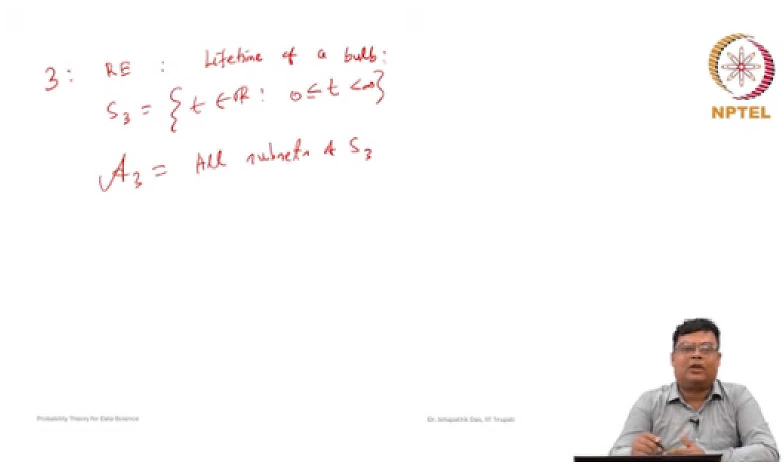


In that case, what is the sample space? This is all t belonging to the real numbers such that t is greater than or equal to 0 and less than infinity. It is a uncountable set. In that case, if you consider the power set, which means all subsets of S3, there will be a problem. So, researchers or scientific communities found that when they consider all subsets of S3, which is the power set, and treat these subsets as events, they face problems defining probability.

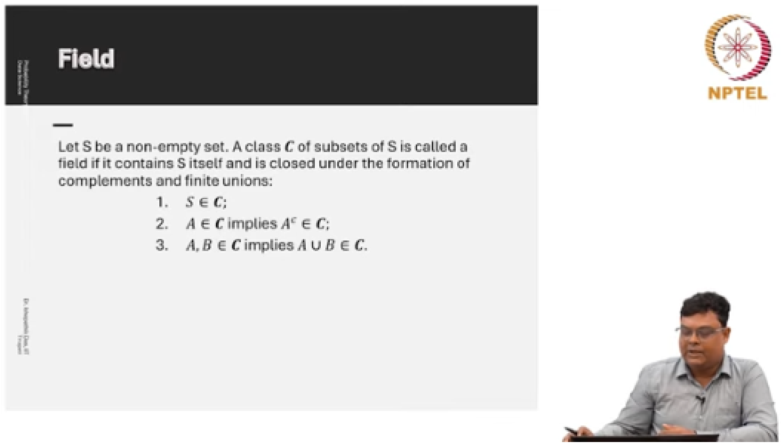
So, in a broader sense, in advanced probability, you may learn that this is called measure theory. So, in measure theory, if you define a measure, you can define the probability on these subsets or events. For countable or discrete sets, it is okay to consider the whole power set. However, for uncountable sets, this approach can be problematic. But if it is an uncountable set, we cannot consider the power set.

If you consider the power set, then you cannot define the probability. It will not be well-defined. So, there will be some contradiction, which can be shown, but we will not go into that direction. We will just understand that when the set is not discrete but continuous or uncountable, we need to define some special subsets, not all subsets. We have to consider a class of subsets of S, the sample space.

That is known as a sigma-algebra. So, all the elements inside that sigma field are known as events, not all subsets. So, first, we will discuss the definition of a field. What is the definition of a field? Let S be a non-empty set.

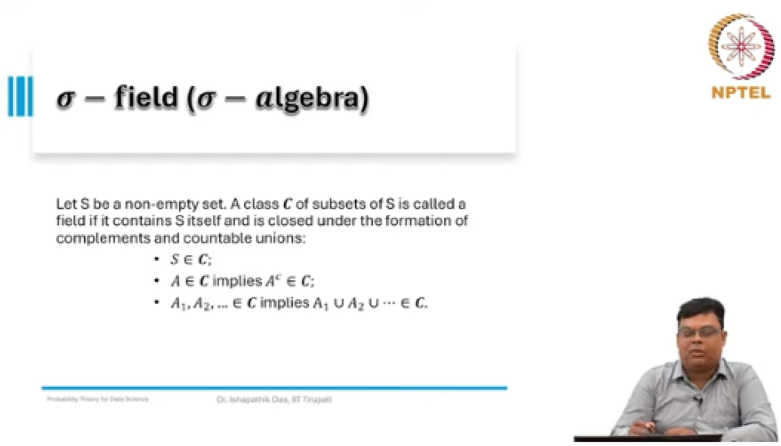


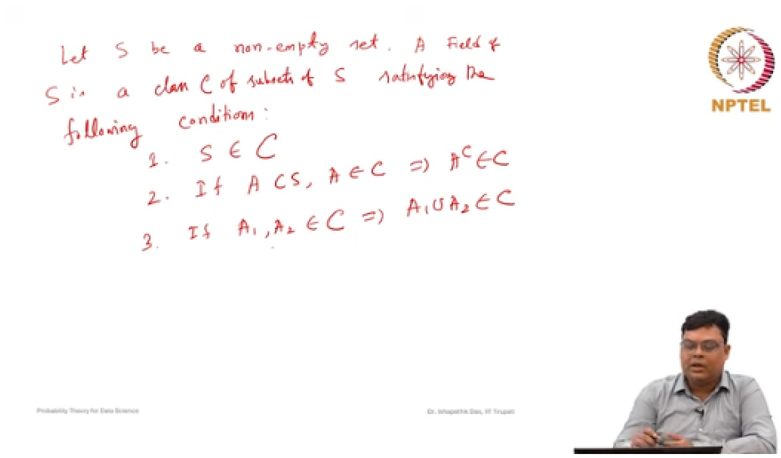
A class C of subsets of S is called a field if it contains S itself and is closed under the formation of complements and finite unions. So, field is defined like that. Let S be a non-empty set. A field of S is a class of subsets of S. This means it is a collection, so it is a class of subsets of S. A field of S is a class, referred to as C, which is a subset of S and satisfies the following conditions.



So, what are those conditions? The first condition is that $S \in C$. The second condition is that if $A \subseteq S$ such that $A \in C$, then A' (the complement of A) also belongs to C. The third condition is that if A_1 and A_2 are subsets that both belong to C, then $A_1 \cup A_2$ (the union of A_1 and A_2) also belongs to C. This means that finite unions are included, and this is known as a field.

It is important to understand the concept of a sigma field. A sigma field is essentially a field with an additional condition. Specifically, a sigma field must be a field and must also satisfy one more requirement. This requirement states that while a field considers finite unions, a sigma field considers countable unions. So, we will discuss some examples.

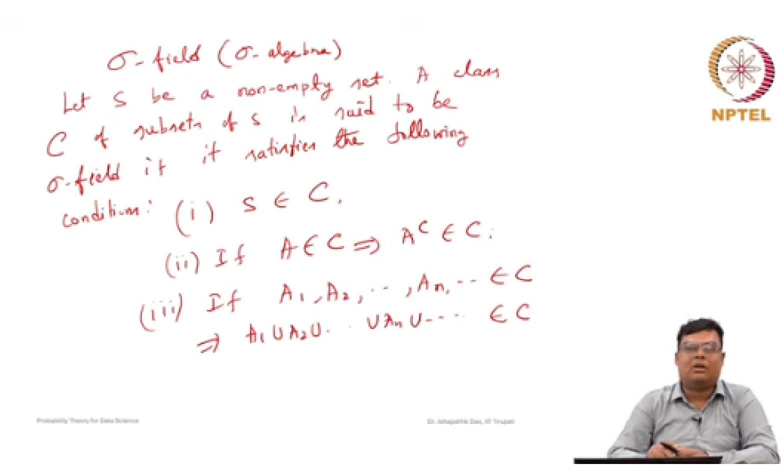




So, what is the sigma field? Sometimes it is known as sigma-algebra. Some books mention sigma algebra, while in others, it is discussed as a sigma field. A sigma field is defined as follows: Let **S** be a non-empty set, and let **C** be a collection of subsets of **S**. **C** is said to be a sigma field if it satisfies the following conditions:

1. **S** ∈ **C**;
2. If **A** ∈ **C**, then the complement of **A**, denoted as **A'**, also belongs to **C**;
3. If a countable collection of subsets belongs to **C**, then the union of these subsets also belongs to **C**.

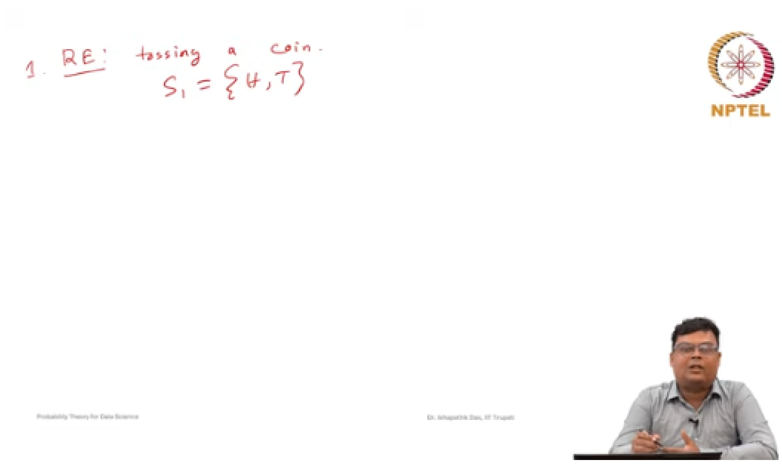
So, this means that if there is a countably infinite collection of subsets of **S** that belongs to **C**, then their union also belongs to **C**. This is known as a sigma field or sigma algebra. Note that if a collection of subsets of **S** is a sigma field, it must also be a field, because any sigma field includes finite unions.



For example, if $\mathbf{A}_1$ and $\mathbf{A}_2$ belong to **C**, then their union, denoted as $\mathbf{A}_1 \cup \mathbf{A}_2$, also belongs to **C**. Since a sigma field is closed under countable unions, it satisfies the properties of a field. This refers to an infinite collection, such as a countable infinite series of subsets like $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, and so on. According to the third property of a sigma field, if each subset in this collection belongs to **C**, then their countable union must also belong to **C**. Therefore, because a sigma field is closed under countable unions, it will also include finite unions. Hence, a sigma field must be a field.

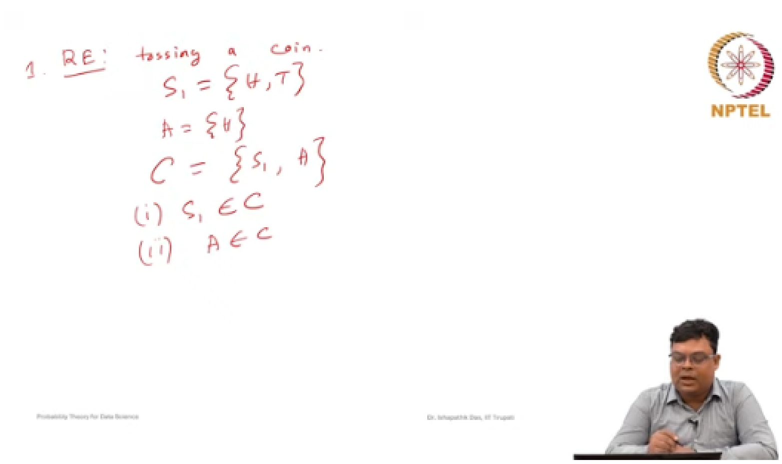
Let's discuss some examples. For instance, consider a random experiment where the sample space $\mathbf{S}_1$ consists of the outcomes "head" and "tail" from tossing a coin. We need to determine if the power set of $\mathbf{S}_1$ is a sigma field. Because the power set contains all the elements, the complement of **S** also belongs to the power set. Any subset you consider will have its complement as a subset of **S**, and that complement will be inside the power set of **S**.

Now, if you consider some countable elements of the power set, those are subsets of **S**. If you take the union of all these subsets, it will also be inside **S**. Therefore, the power set will be a sigma field. However, we are not interested in the power set because if **S** is an uncountable set, then the power set will face difficulties in defining the probability. So, let us define some sigma fields that are not power sets.



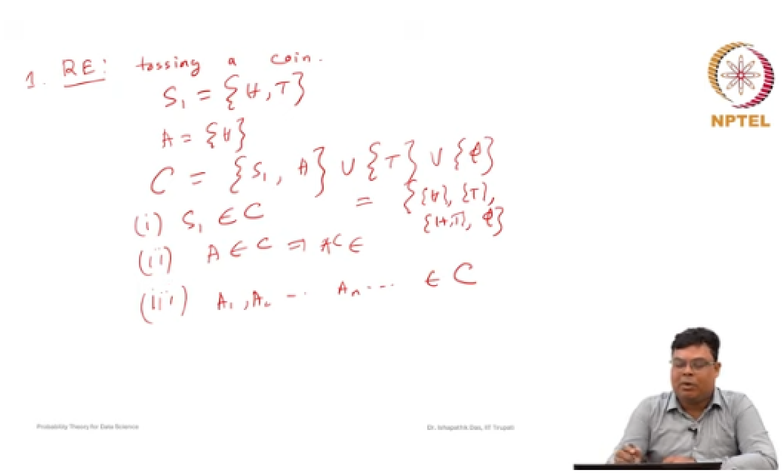
So, let us consider a case where **A** is just **H**. Now, I want to check whether the collection containing $\mathbf{S}_1$ and **A** will be a sigma field. See that $\mathbf{S}_1$ belongs to the collection, so the first condition is satisfied: $\mathbf{S}_1 \in \mathbf{C}$. The second condition is that if $\mathbf{A} \in \mathbf{C}$, then its complement must also belong to **C**.

However, the complement of **A** is "tail," which is not in **C**. Therefore, it is not a sigma field. Hence, to make it a sigma field, we will add some of the sets inside. So, let us take the union of the complement of **A**, which is "tail." Including the union of "tail" as well, our set will also include $\mathbf{S}_1$, the complement, and the empty set $\emptyset$.



Thus, C will become the power set. So this is nothing but head, tail, S1, and the empty set. So, if you include all these sets, you can see that it forms a sigma field. Because A belongs to C, its complement, which is tail, also belongs to C. Since there are only four elements, we need to check only finite unions.

You can include the empty set as well. For any infinite collection, if you add sets like the empty set, you can see that this will be in C. Thus, this will be a sigma field, but it is again the power set. But what we did is this: whenever it is not a sigma field, we add some sets to make it a sigma field. So, we add subsets of S to make it a sigma field.



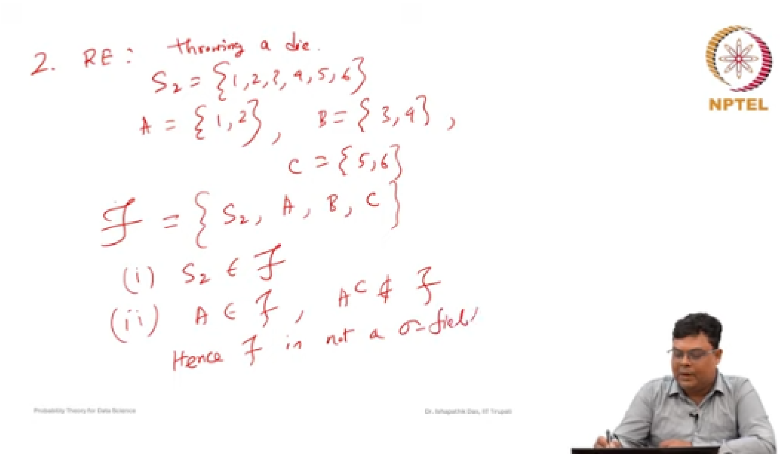
This kind of process is known as the sigma field generated by the set. So, we will take a collection. Let us consider a second example involving a random experiment. Suppose this is $\mathbf{S}_2$. In this case, $\mathbf{S}_2$ contains {1, 2, 3, 4, 5, 6}. Now, let us consider **A** as {1, 2}, **B** as {3, 4}, and **C** as {5, 6}.

Suppose we take this collection where **C** is present. Let's use a different notation: suppose **F** is the class of subsets of $\mathbf{S}_2$ containing $\mathbf{S}_2$ itself, **A**, **B**, and **C**. Let us determine whether this is a sigma field.

Condition one: $\mathbf{S}_2 \in \mathbf{F}$. Condition two: **A**, which is {1, 2}, must also be in **F**. So, the complement of **A** will be {3, 4, 5, 6}. This is not in **F**.

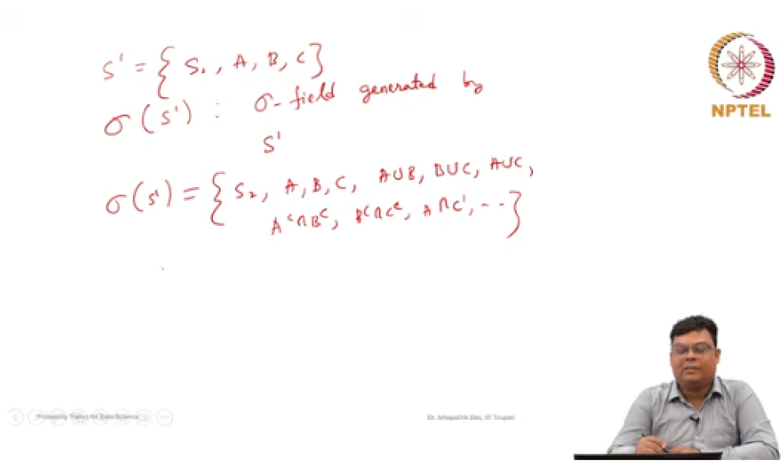
So, what will we do? We will add this set. So, this is not a sigma field. First of all, suppose **A** is {1, 2}. $\mathbf{A} \in \mathbf{F}$, but we see that the complement of **A** does not belong to this collection **F**. Hence, **F** is not a sigma field.

F is not a sigma field. So, then what will we do? We denote that, suppose, we say this is a set **S'** here. **S** is the set containing all these collections, such as **A** and $\mathbf{S}_2$. Let us use another notation here.



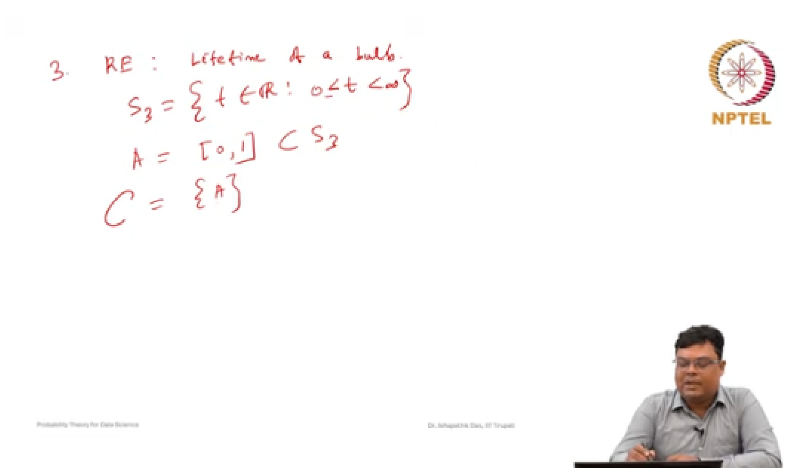
Let **S'** represent the set containing **A**, **B**, and **C**. This set is not a sigma field. We denote this as the sigma field generated by **S'**. So, what is that? You include many more other sets that are required. So, what will be the sigma field generated by **S'**?

Now, $\mathbf{S}_2$, **A**, **B**, **C**, their unions, $\mathbf{B} \cup \mathbf{C}$, $\mathbf{A} \cup \mathbf{C}$, and their complements are also included. This means $\mathbf{A'} \cap \mathbf{B'}$, because by De Morgan's law, the complement of $\mathbf{A} \cup \mathbf{B}$ is $\mathbf{A'} \cap \mathbf{B'}$, and similarly $\mathbf{B'} \cap \mathbf{C'}$, $\mathbf{A'} \cap \mathbf{C'}$, and so on. You include many more sets, and then if you take the union, such as $\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}$, which is $\mathbf{S}_2$, you will see that it will be a sigma field. So, this is called the sigma field generated by the set **S**. Now, let us consider the third example.



Now, let us consider the third example. This random experiment involves measuring the lifetime of a bulb. So, for the lifetime of a bulb, if you consider it, what will be our sample space? This is t, which belongs to the real numbers such that t is greater than 0 and less than infinity. Now, let us consider one set, which is 0 to 1. Now, we want to find the sigma field.

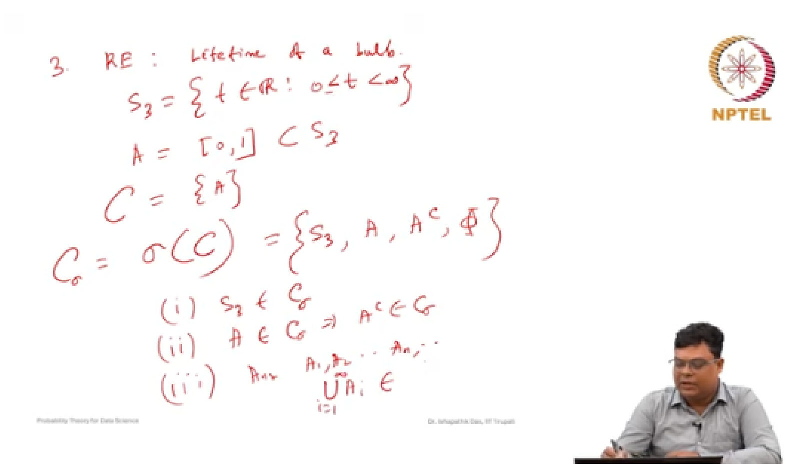
If we consider only this set, suppose your C contains only this subset. So, it is a subset of S3, obviously. Now, if C is equal to A and contains only the set A, will it be a sigma field? It will not be a sigma field because A complement is nothing but the set from 1 to infinity. It will not be a sigma field.



We will consider a sigma field generated by C. So, this is C. This is the sigma field generated by C. Let us denote it by Cσ. What will be the set? Since it has to be a sigma field, S_3 has to be included here. A is already there. Therefore, the complement also has to be included, and the null set is there as well. Now, you can see that this is a sigma field. You can add more sets also so that it will be a sigma field. But we want to add that many sets, which is becoming just a sigma field.

So, it is called the smallest sigma field generated by the set C. So it will be a sigma field because if you consider S_3, it belongs to $C\sigma$. The second condition is that if A belongs to $C\sigma$, then the complement of A also belongs to $C\sigma$. This is because the complement of the complement of A is A. If you take the union, this is nothing but S_3.

If you take the intersection, this is nothing but the empty set, $\emptyset$. All these belong to the sigma field. So, any collection also, if you consider, because there are only a finite number of elements, that is why for any collection A_1, A_2, ..., A_n, if you need to make it infinite, you have to include the empty set and the sets A_1, A_2, etc. Taking the union of A_i where i ranges from 1 to infinity, this countable collection will also be inside C. So, this will be a sigma field.



So, this is some kind of concept of a sigma field we discussed. So, this is required whenever we define probability. This sigma field concept will be required, that's why we discussed it. You may go through in detail some of the books also we refer, so there you can find more details about this sigma field and some of the examples we discussed. So, please go through it and see whether you have understood and get clarity.

Next, we will discuss how these events are defined because this is a subset of a sample space. Some of the set operations we need to know are very important, such as union and intersection. Maybe you have already learned those things.