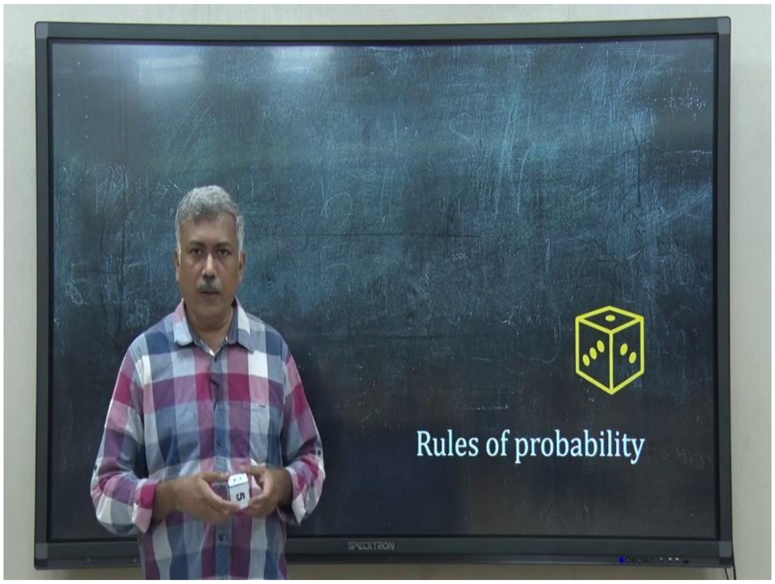
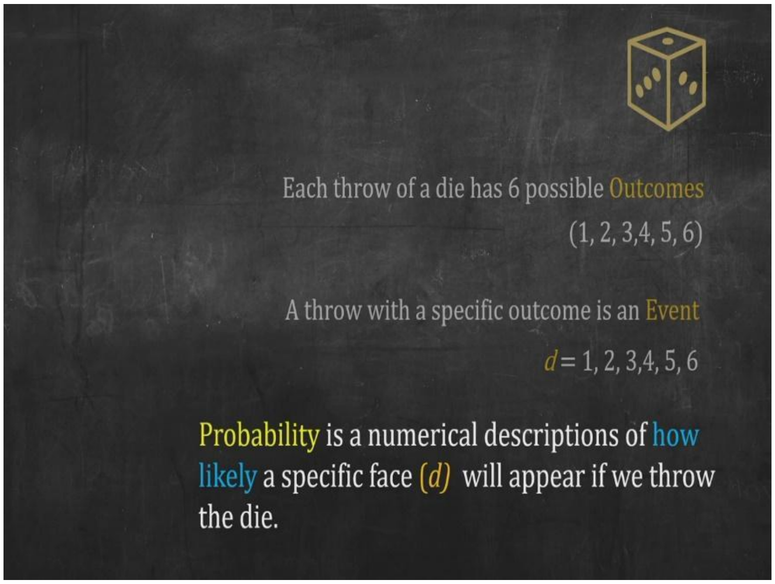
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture: 1
Rules of Probability

(Refer Slide Time: 00:29)

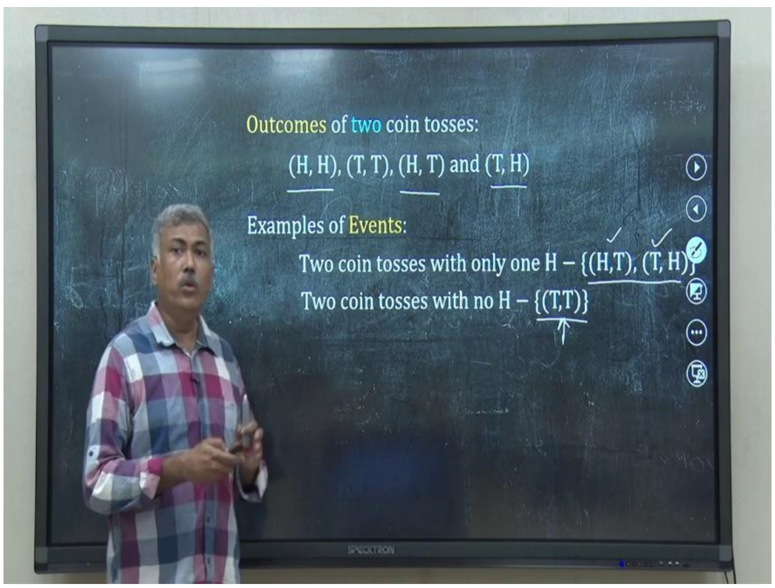




Hello, welcome to our online course on Data Analysis for Biologist. This is the first lecture. In this lecture we will discuss about some basic rules of probability. Now, any discussion on probability cannot be started without discussing dice throw and you know, a coin toss.

So, I have a die in hand. Now, if I throw this I can have 6 outcomes right. I can have 6 outcomes which are 1, 2, 3 or 4, 5 and 6. Now, these all are the outcomes of a die throw. Now, if I throw it and I get suppose 6, right? So, that is one event. So, now, I can ask what is the probability that if I throw this die, I will get a particular face or particular value, 6, 1, 2 or something like that. Now, the probability, by probability what we mean is that you need. I want a numerical description of likelihood that a specific face of this die will appear when I throw it. Now, the idea of outcomes and events are sometimes a bit difficult to understand. So, let us take another example.

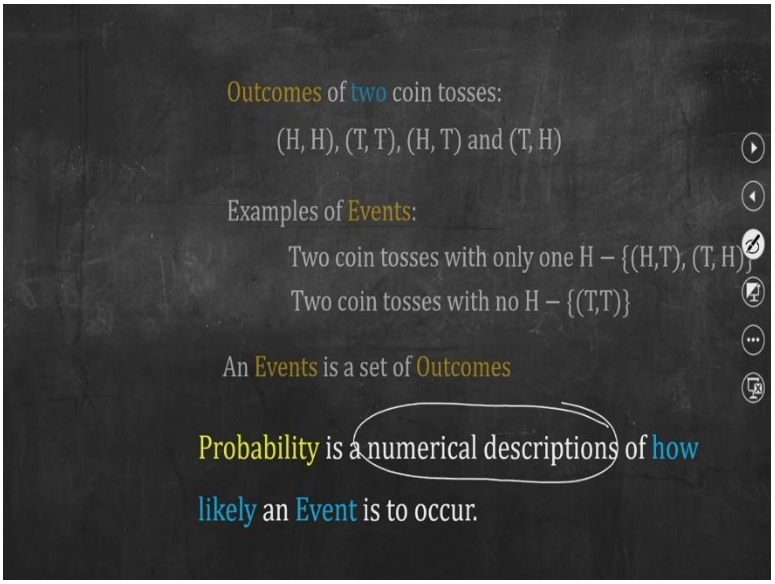
(Refer Slide Time: 01:38)



Suppose I have a coin or rather suppose I have two coins here, and I throw them, I toss them. So, each of them if I toss I will get a head or a tail. Now, if I toss both of them one after another, then I can have multiple results right, multiple outcomes. For example, one outcome could be that you get H, H, or one outcome could be T, H, tail and head. The other outcome could be H and Tail. So, I have four possible outcomes.

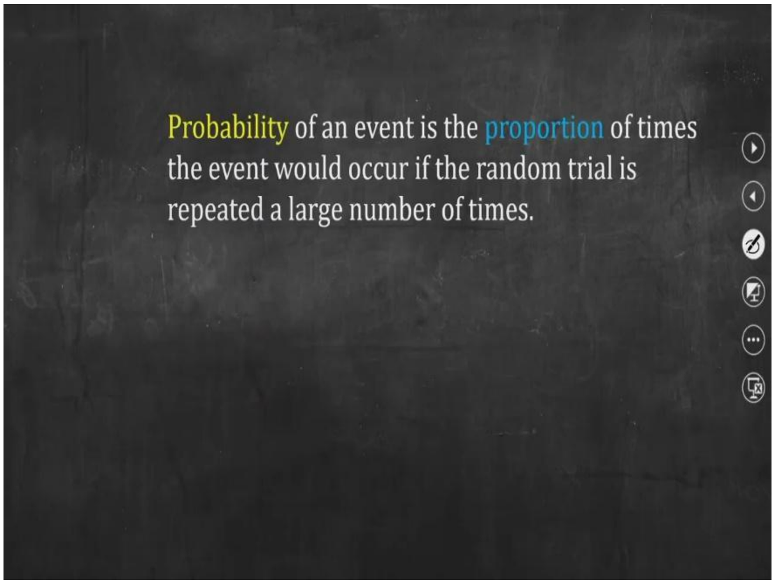
Now, suppose if I define that okay, I will toss these two coins to one after another. And then I want to know that okay, whether I have got one H out of this, one head out of this coin toss or two-coin tosses or not. So, that is one event right. So, I can get this one head out of two coin tosses by two way. One is H and T, the other one is T and H or suppose, I have tossed it twice and I want that, no H has appeared. Then that can happen only in one way that is T, T. In both cases, I have got tail and tail. So, in this case, what I can ask, tell me, what is the probability that if I toss the coin twice, I will get no head? So, I want to know the probability of this second event that no H will appear.

(Refer Slide Time: 03:14)



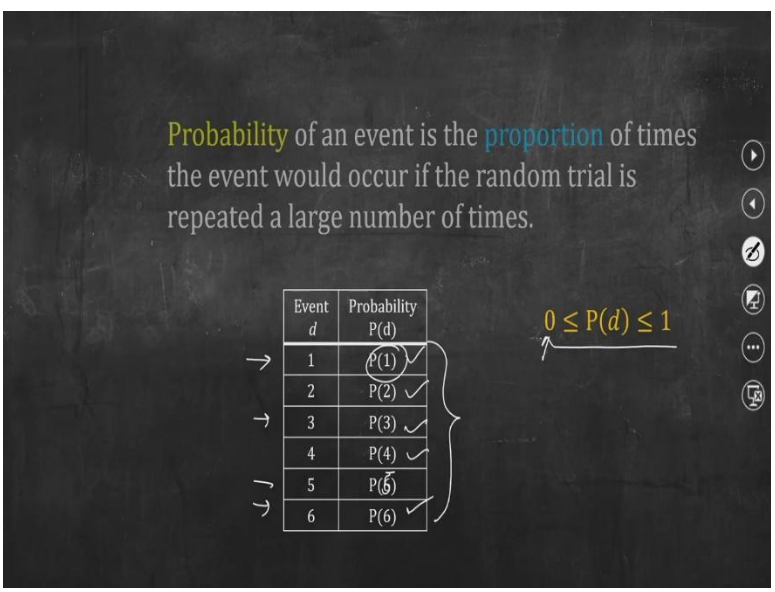
So, again, what I am asking is that, I am asking you to give me some numerical description of how likely that no H, no head will appear, if I toss the coin twice. So, then I can generalize it further, I can say probability is a numerical description of how likely an event is to occur. Now, what do I mean by this numerical description right? That is one important question. I have to understand what do I mean by this numerical description? By numerical description, what we essentially means that we want some sort of relative frequency. What do I mean by relative frequency?

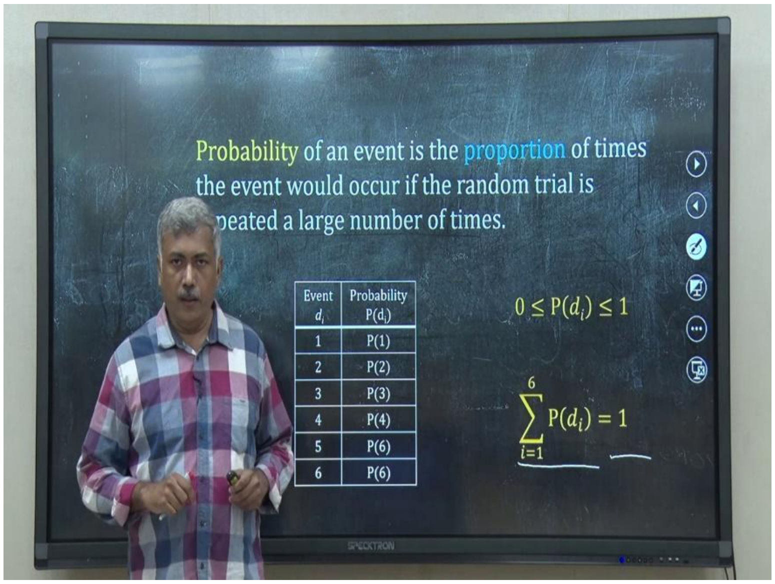
(Refer Slide Time: 03:55)



Suppose I repeat the trial, I repeat the coin toss, I repeat the throw of dice multiple times, large number of times, hundreds of times. And then I asked calculate the proportion of times when that particular event, like I have tossed twice 100 times, and I have not got any hit in each of those two coin tosses. So, those are the events. And I am asking you, what is the proportion of times that, that event has happened when I have random done random trials for a large number of times.

(Refer Slide Time: 04:26)





So, now take the example of a die throw. In die through I have 6 possible event I can have right, I can get 1, if I throw it, I can get 3, 2, 5 and 6. Now suppose each of these has some associated probability right. Each of them has some associated probability. I may have calculated them by a large number of throw of die or some other way.

I know this probability and I represent those probability as P(1), P(2), P(3), P(4) and P(5) and P(6). So, as these are relative frequencies, as these probabilities are essentially proportion, then what I can say that each of these P, these each of these probability should be less equal to 1, and they will be positive. So, they should be greater equal to 0.

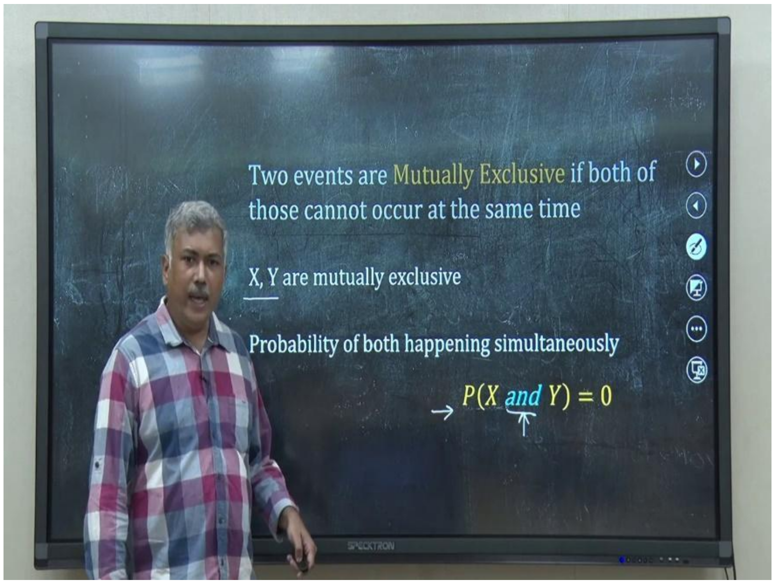
$$0 \leq P(d_i) \leq 1$$

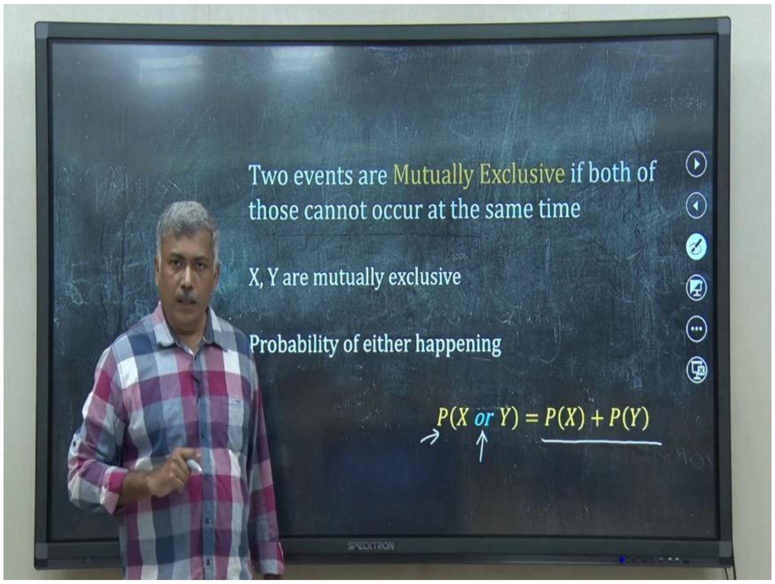
Now, these are all relative frequencies, probabilities are proportions relative frequency. So, if I sum them all this probability, if I sum all this probability, then what do, I get, I should get,

$$\Sigma P(d_i) = 1$$

Now, when we discuss about probability and statistics, we have to remember two very important terms, one is called Mutually Exclusive Event and Independent Events.

(Refer Slide Time: 05:49)





So, what is Mutually Exclusive Event? If I have two events, like for example, I am throwing the die and I get 1 or 2 or 3 or 5 or 6, something like that right, each of them is event. So, now two events are mutually exclusive when two of those cannot happen simultaneously at the same time. If I toss a coin, I will either get head or tail, both head and tail cannot happen together.

So, these are called Mutually Exclusive Event. Either a DNA base is mutated or it is not mutated, either a bacteria has got killed or it has not got killed. So, these are all mutually exclusive. Now, if I have two random variable events, which are X and Y. And they are Mutually Exclusive Event.

Then what I can say that as they are mutually exclusive, the probability that both of them will happen simultaneously is equal to 0. So, how do I represent that? I represent that as,

$$P(X \text{ and } Y) = 0$$

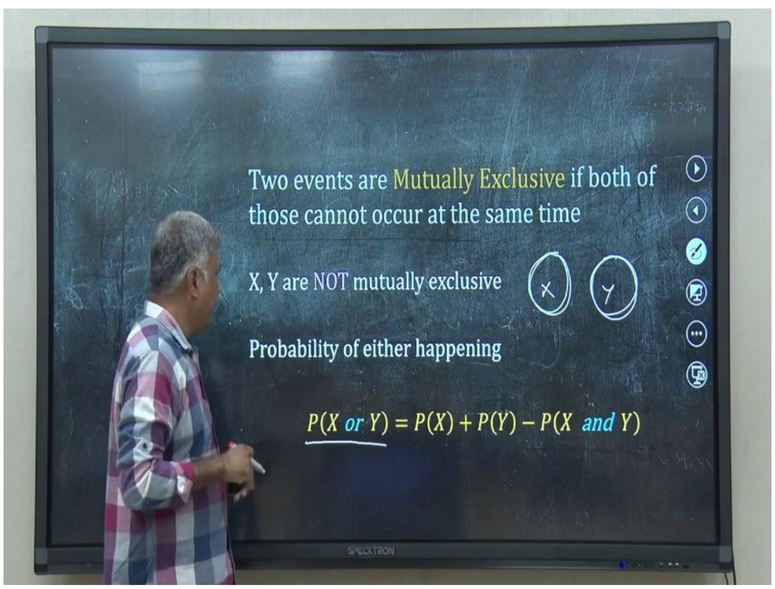
Notice this AND, that means both of them are happening together, and that should be equal to 0. Now, if X and Y are Mutually Exclusive Event.

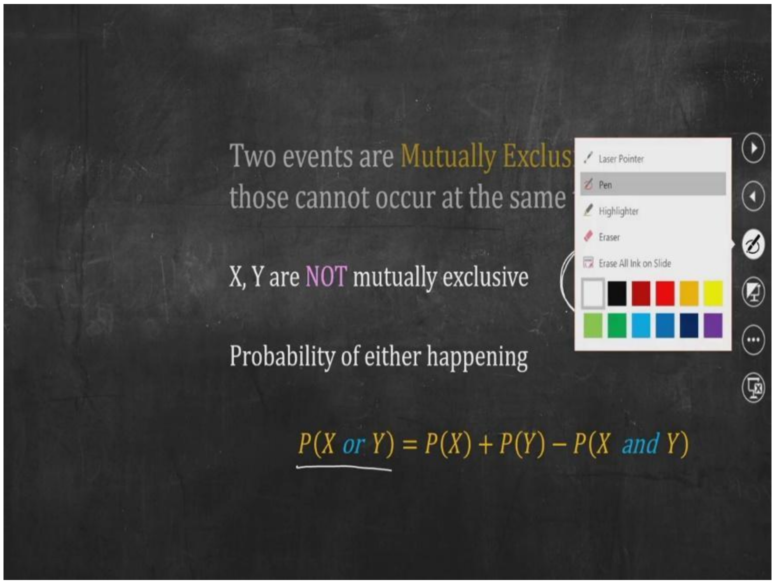
Then what if I asked what is the probability that either of X or Y would happen? So, that would be P(X or Y). Notice here and is replaced by or, so, either X or Y will happen, I want to know the probability of that.

$$P(X \text{ or } Y) = P(X) + P(Y)$$

So, probability of X or Y would be equal to probability of X plus probability of Y.

(Refer Slide Time: 07:30)

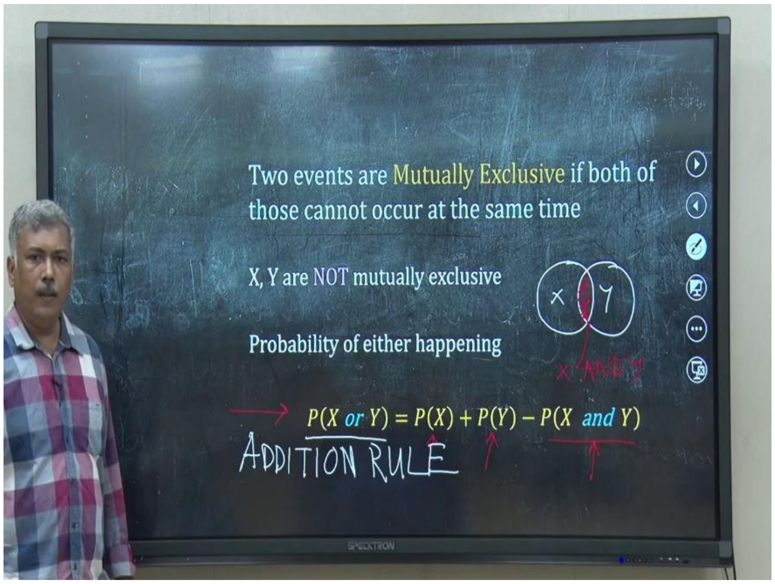




Now, what if X and Y, are not mutually exclusive? So, in that case, what will be the probability that either X or Y will happen. So, again, I want to calculate the probability that X or Y, but X and Y are not Mutually Exclusive. So, what will be that probability. To understand that let us use some graphical method.

So, suppose I consider X and Y are mutually exclusive. and these circles that I have drawn here roughly represent their whole probability space. then, as they are mutually exclusive, they should be completely separated. Now, suppose they are not mutually exclusive. That means, there should be some overlap between these two circles. Let me erase those and draw again.

(Refer Slide Time: 08:24)



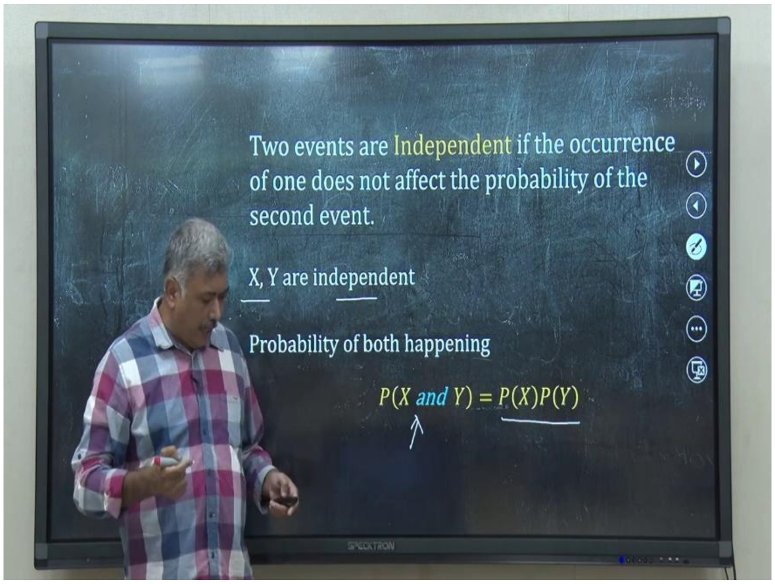
So, if they are not mutually exclusive, then X and Y circles that I have drawn earlier will have, some overlap. and what is this overlap region? This overlap region is this red color thing that I am drawing. and what is that, that is X and Y, both when X and Y has happened simultaneously. So, to get the probability that X or Y will happen, what I have to do?

I have to take probability of X plus probability of Y and then I have to subtract that shared area otherwise, I am counting the same thing twice. So, probability of X or Y when X and Y are not mutually exclusive is,

$$P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y)$$

Now, this rule is known as the ADDITION RULE. So, we call it ADDITION RULE. Now, there is another important point, apart from Mutually Exclusive Event, events can be mutually independent. Let us see, what is that?

(Refer Slide Time: 09:52)



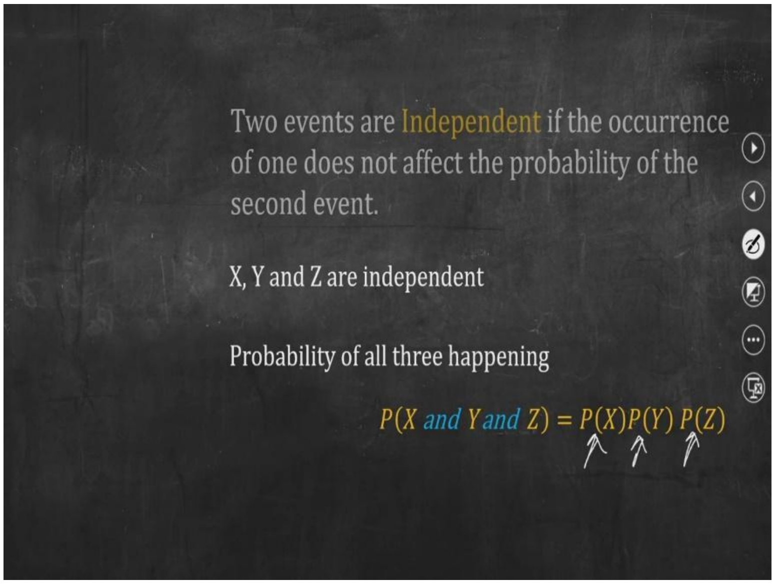
Two events are mutually independent if occurrence of one event does not affect the probability of the other event. For example, you are suppose you are studying the mistakes in DNA sequencing. So, one error at one particular base of the DNA sequence does not affect the probability of another error somewhere else.

So, we can consider these two errors in your DNA sequencing results are mutually independent. So, now if two events are independent, then I can ask okay, X and Y are mutually independent, then I ask what is the probability that both will happen at the same time. So, then I am asking you to calculate the probability of X and Y because both will happen. And that would be equal to,

$$P(X \text{ and } Y) = P(X)P(Y)$$

because they are independent.

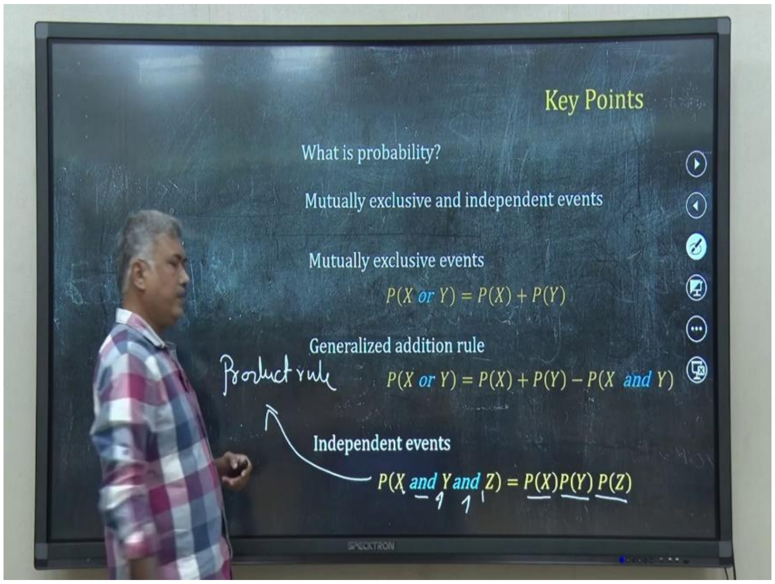
(Refer Slide Time: 10:51)



Now, suppose I have another third event X, Y and Z. So, in that case, probability that X and Y and Z, all will happen is equal to, as they're independent,

$$P(X \text{ and } Y \text{ and } Z) = P(X)P(Y)P(Z)$$

(Refer Slide Time: 11:08)



Now, let us jot down what we have learned in this particular lecture. I have discussed what is probability? Probability is a numerical measure of likely, how likely a particular event can happen. and that numerical measure is nothing but some sort of proportion, relative frequency of how often that event can happen if we do large number of trial or experiments.

Now events can be mutually exclusive, or independent. Now based on this mutually exclusiveness, you know, if I have two mutually exclusive events, X and Y, then probability that either of them will happen,

$$P(X \text{ or } Y) = P(X) + P(Y)$$

and this is called the ADDITION RULE.

We can generalize this addition rule if, considering that X and Y are not mutually exclusive, in that case,

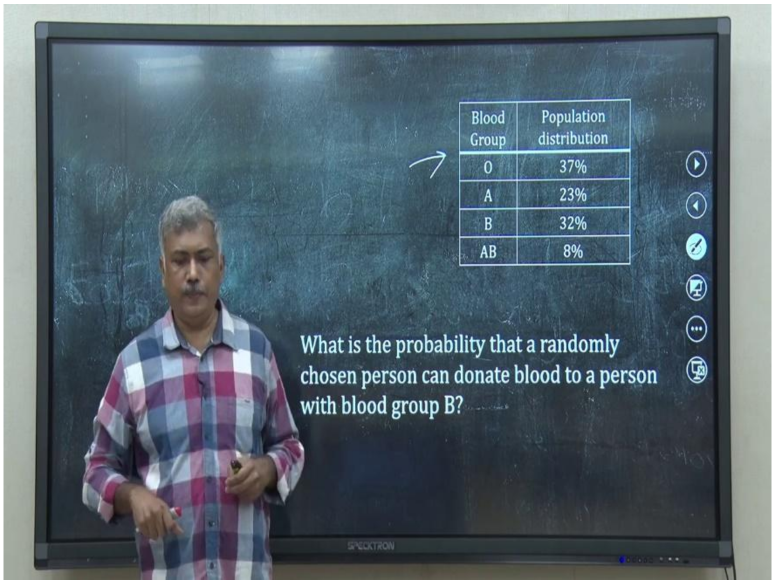
$$P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y)$$

Now, if X and Y are independent event, then we get the product rule. What is it, in that case,

$$P(X \text{ and } Y \text{ and } Z) = P(X)P(Y)P(Z)$$

and this is what we call product rule. With this, I will end this lecture. Before we end, let me give you a problem to solve..

(Refer Slide Time: 12:46)



Suppose we have collected data from population. and we have tried to identify the population distribution of four blood group O, A, B and AB, and that data is given here. Now, the question is if I pick someone randomly from the population, without any bias, you are randomly picking someone from the population. and then if I ask, can that random person donate blood to a person whose blood group is B?

So, what I am asking you to calculate using the data given here, what is the probability that a randomly chosen person can donate blood to a person with blood group B? To answer this question, you have to use either the rule of addition, product rule or the idea of mutually exclusive and the idea of mutually independent that we have discussed in this lecture. So, try to solve this one, till then happy learning. Thank you.