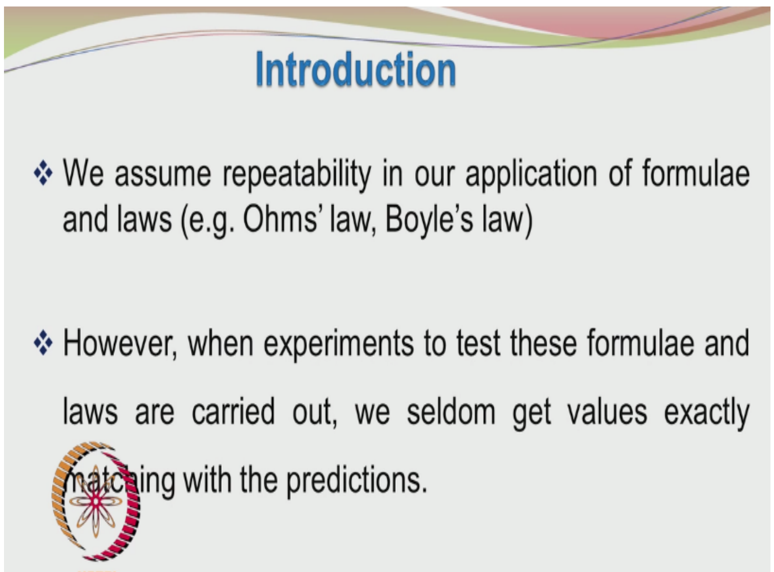
Statistics for Experimentalists
Prof. Kannan A
Department of Chemical Engineering
Indian Institute of Technology – Madras

Lecture – 02
Random Variables

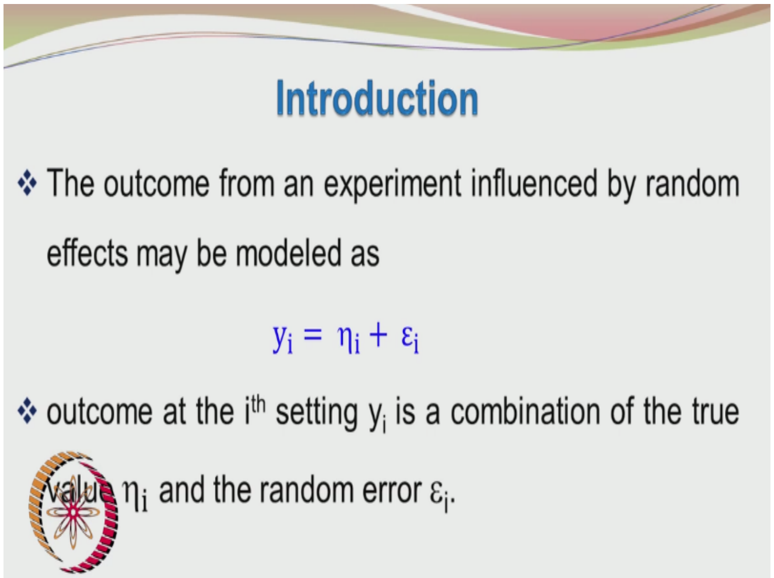
Now we will be looking at random variables. The reference book for this particular section is the one written by Ogunnaike. The title of the book is Random Phenomena. It is a fairly reason book.

(Refer Slide Time: 00:36)



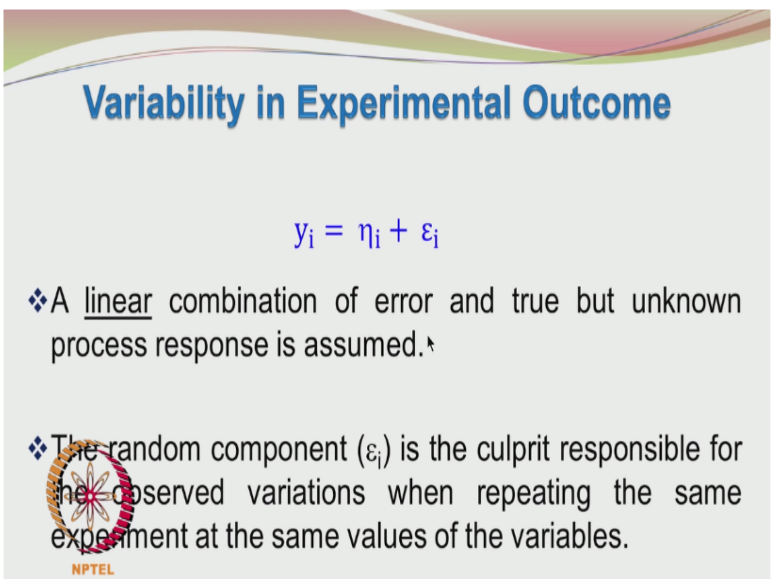
We know Ohms' law and Boyle's law. A law is something which cannot be questioned. So when we carry out experiments in accordance with these laws we expect that all the experiments should be giving the same value as specified by the law. However, when we do these experiments in order to verify the formulae or demonstrate these formulae in the laboratory, we seldom get values exactly matching with the predictions.

(Refer Slide Time: 01:30)



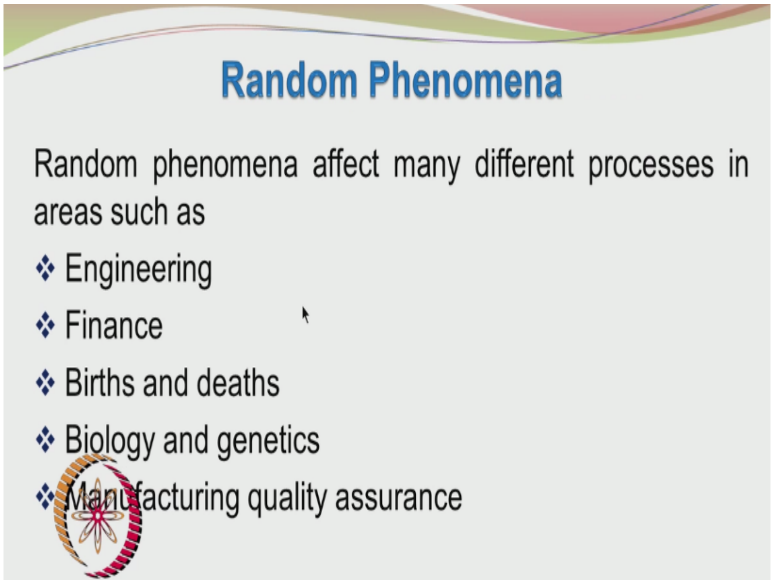
What happens is, you have the true value here, and the actually observed value here. The 2 are not identical. You can see that you also have an extra term epsilon i in this equation. This is the random error component. It can take a positive value or it can take a negative value. So the response may be higher than the value given by the law or lower than the value given by the law depending upon whether the error component was positive or negative.

(Refer Slide Time: 02:25)



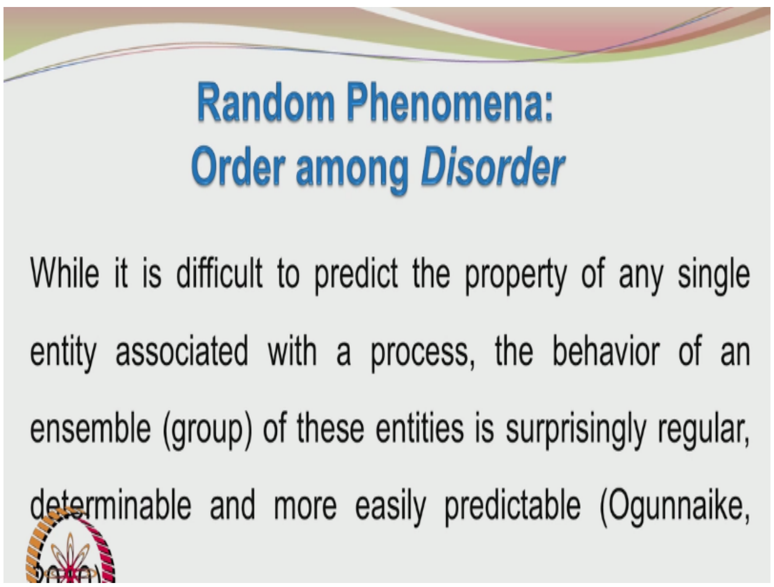
What we have also implicitly assumed is the combination of the true value and the error is a linear one. We are putting epsilon i added on to eta i a linear combination. We did not put eta I * exponential of 1 + epsilon i or exponential of epsilon i or we did not put sin of epsilon i. We did not have a nonlinear combination. We are having a linear combination, simple addition of the error term to the expected response. This random component epsilon i is the main reason for the variations we observe in the experiment.

(Refer Slide Time: 03:30)



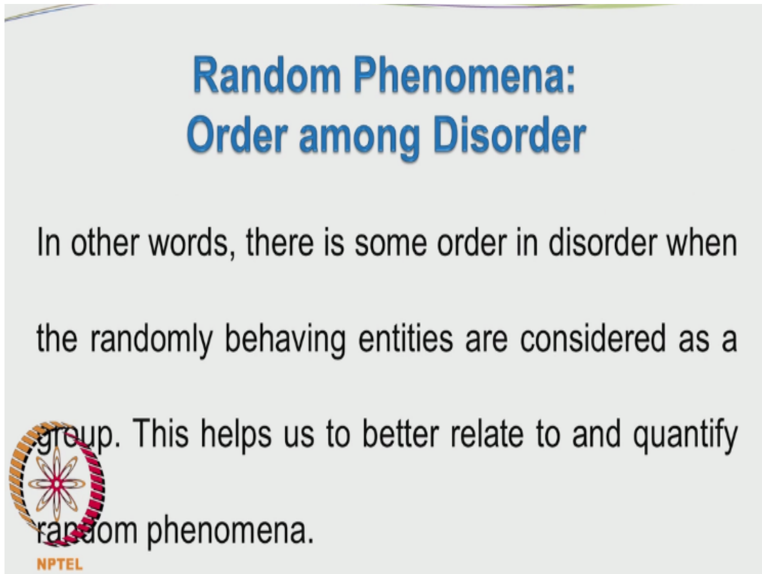
So we call all the variable phenomena as random phenomena and we encounter random phenomena in different walks of life. We may encounter them in engineering, finance, births and deaths, biology and genetics, manufacturing quality assurance and so on.

(Refer Slide Time: 04:02)



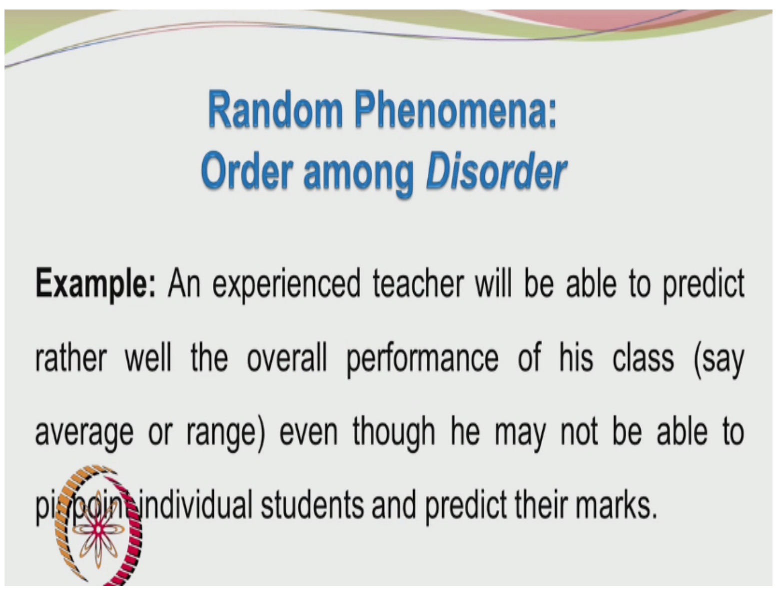
People who are been working in this field they came up with some interesting observations. They found order among disorder. They found that an individual entity is difficult to predict, but a collection of such individual entities behave in more or less a predictable manner. A group of entities is more easy to predict than a single entity in the group. This is a very interesting observation and if you reflect on it, you may find many examples. I will give a simple on shortly.

(Refer Slide Time: 05:08)



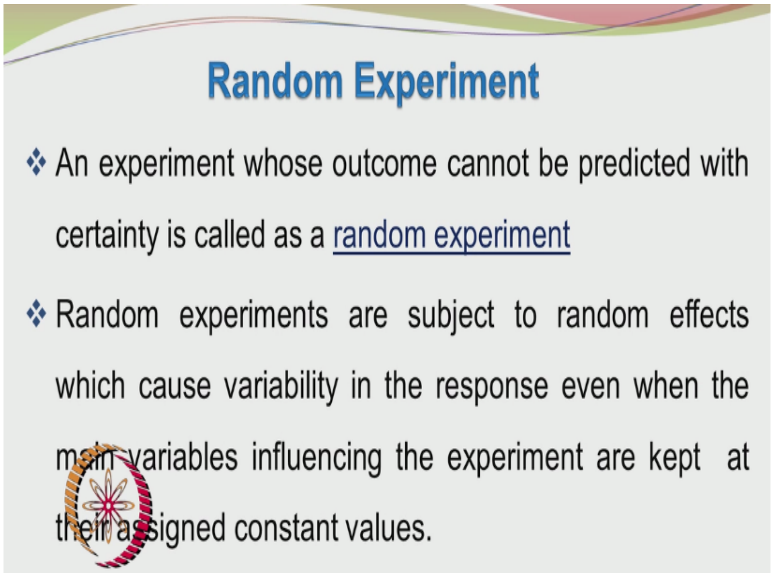
So there is some order in the so called disordered entities or random entities. So this really helps us to better relate to random phenomena and also to quantify the random phenomena.

(Refer Slide Time: 05:31)



A simple example would be an experience teacher in a class. He will be able to predict rather well the overall performance of his class in terms of average or range even though he may not be able to pinpoint the individual student's marks. He may say that the highest marks is likely to be about 80 and about 10% of the class will get between 70 to 80 and may be 5% of the class will get marks below 30. So he will be able to better relate to a class of students.

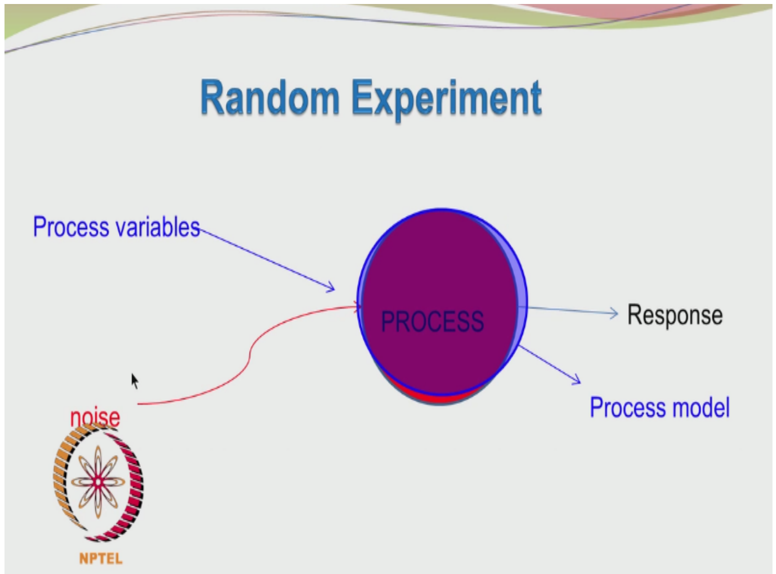
(Refer Slide Time: 06:26)



So now we come to the random experiment. Random experiment is defined as one whose outcome cannot be predicted with certain thing. So when you throw a die any number between 1 to 6 may show up. You cannot certainly say I am throwing the die; I am going to get the number 1. So this is a random experiment. So the random experiments are subject to random effects which cause variability in the response even when the main variables influencing the experiment are kept at their assigned constant values.

For example, even if you control the temperature of the reactor at exactly 90 degree centigrade for example all other conditions like pressure, feet concentration, type of catalyst, everything you are keeping as constant. The output from the reactor may be in terms of yield or conversion or concentration of the desired product may be different for different experiments and so these experiments are called as random experiments and we are not able to predict with certainty the yield from the reactor is going to be this much or the percentage conversion will always be exactly 90% or 70% whatever.

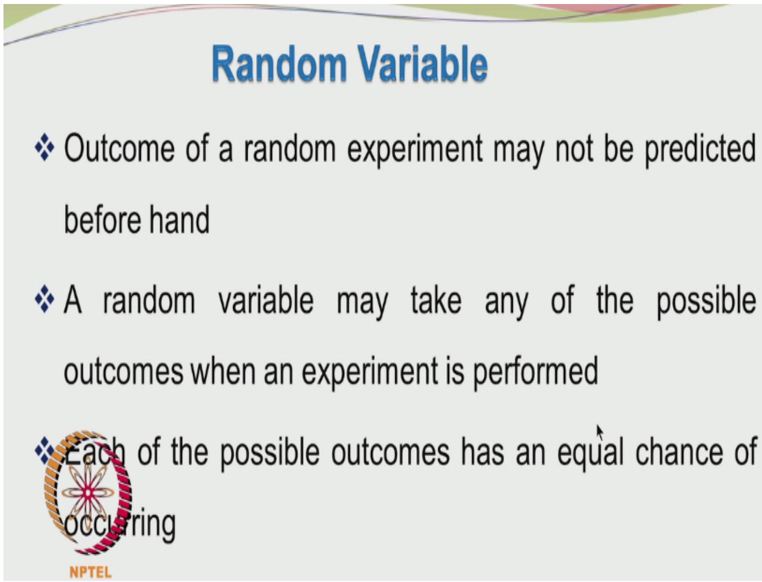
(Refer Slide Time: 08:05)



So this diagram shows the ideal process and the actual process. The ideal process which is one given by the model is shown by the blue diagram and here you have the process variables which are influencing the process. They are the inputs to the process model. You also have the unwelcome inevitable noise affecting the process and so you can even say that because of the effect of the noise which has not been accounted for in your model.

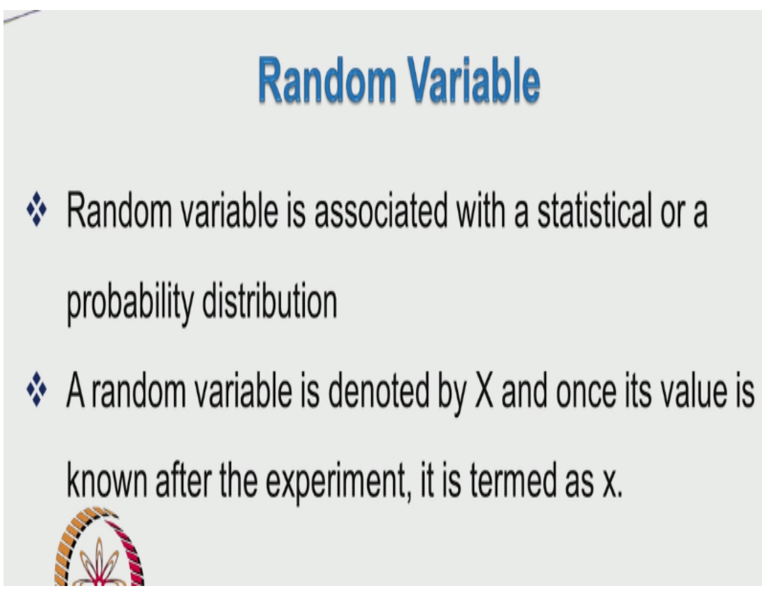
The model gets distorted and becomes the red colored circle which represents the actual process. The blue curve was the model and the red curve is actually the process. so the response from the process is different from the prediction from the process model. The main reason is the process is not exactly as modeled and it does not exactly as model because of the influence of the noise. So our response is different from the process model predictions.

(Refer Slide Time: 10:00)



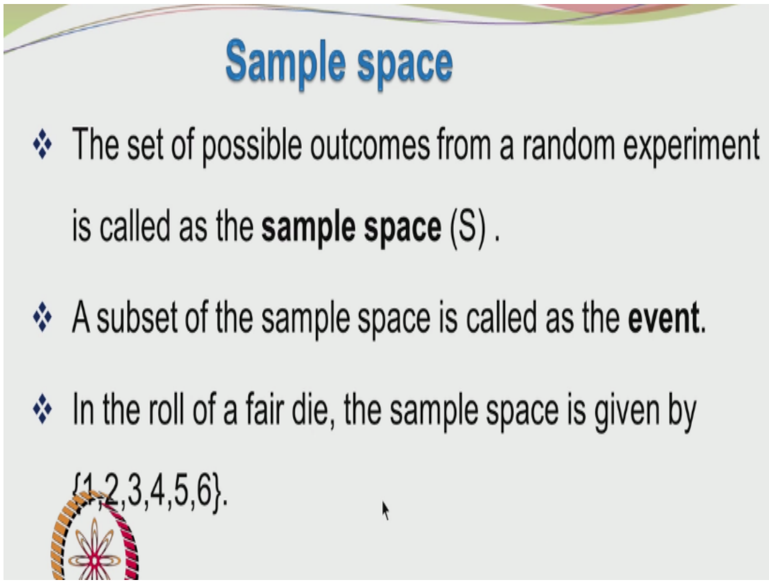
What the characteristic feature of a random variable. The outcome of a random experiment may not be predicted beforehand. Random variable may take any of the possible outcomes when an experiment is performed and each of the possible outcomes has an equal chance of occurring. Random variable is associated with the statistical or a probability distribution. So random variable can take several possible values and these values are associated with the probability distribution. We will be seeing more on this in a short while from now.

(Refer Slide Time: 10:53)



The random variable can take a discrete value or it can have a range of continuously varying values. So we will be seeing those issues very shortly. How do we represent the random variable? The random variable is denoted by x. It is an abstract entity before you perform the experiment. It is just given a place holder or a name as x and once the value is known after the experiment, it is given a terminology of small x. The small x is actually a value.

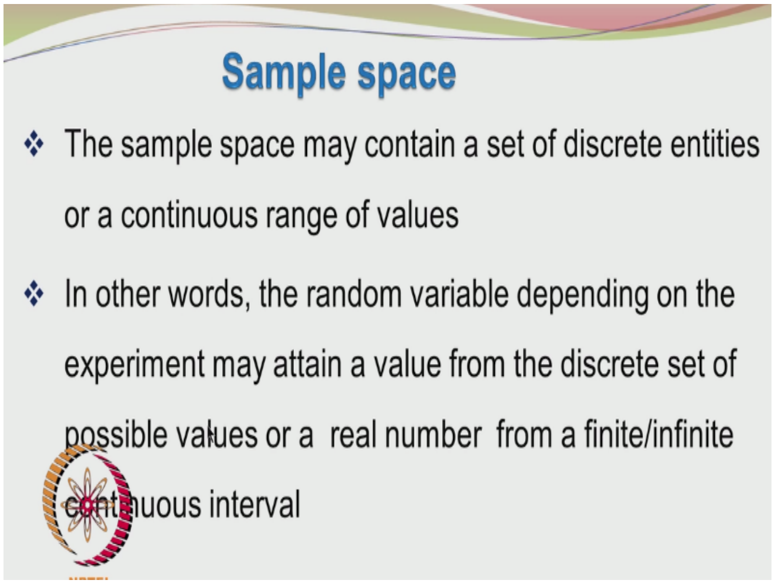
(Refer Slide Time: 11:48)



Now we define sample space. It is the set of all possible outcomes from a random experiment. A sample space contains or includes all the possible outcomes from the random experiment. A subset of the sample space is called as the event. When you roll a fair die, the sample space is given by 1, 2, 3, 4, 5, 6. These are the different possible outcomes of the sample space and a subset of it would be if I roll a fair die.

What is the outcome of number 2, or number 4, or number 6 or generally you can classify it as the event is the occurrence of an even number. So the subset of the sample space in that case would be 2, 4, and 6. You can say what would be the chances of getting a odd number. Then the event would be 1, 3, or 5 and that would be the event which is a subset of the entire sample space. So you cannot have the event taking values outside this sample space.

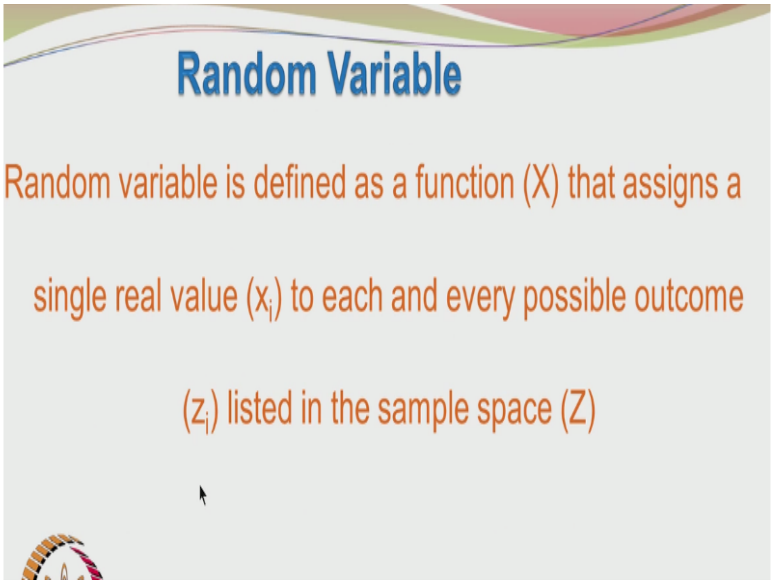
(Refer Slide Time: 13:30)



In the previous case, we saw the sample space taking values between 1 and 6 both included and these were discrete or separate entities and that depended upon the nature of the experiments we performed, but in many experimental work the possible values that may be taken by a random variable need not be discrete, it may be continuous also.

For example, when we are measuring the temperature of a process vessel, there is no need that the temperature will take only discrete values. It can take any value between let us say 0 to 100. We are measuring the temperature of water which is being heated so it can take any value between 0 degrees to 100 degrees provided we are boiling at normal atmospheric pressure.

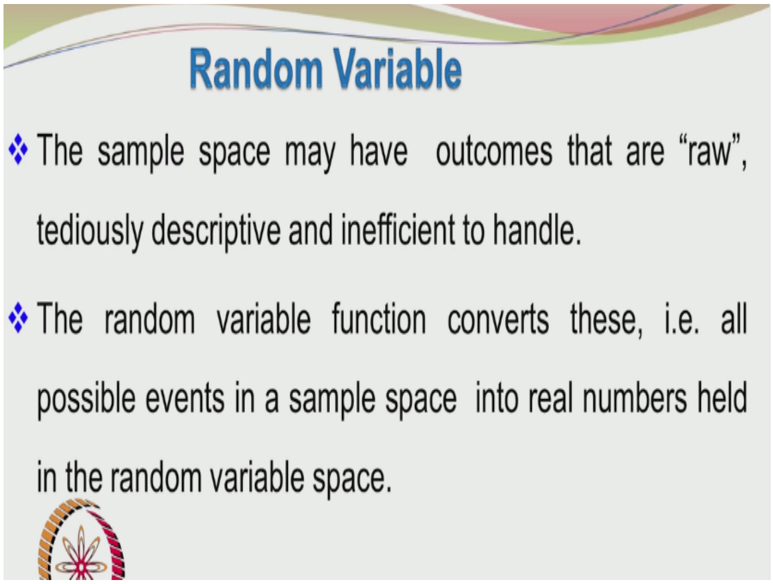
(Refer Slide Time: 14:38)



Now we are going to see another role of the random variable. This is very interesting. We have been so far referring it as a variable and now we are going to refer to it as a function. It may seem a bit confusing, but if you follow the discussion it is in fact quite simple. Random variable is defined as a function that assigns a single real value to each and every possible outcome in the sample space.

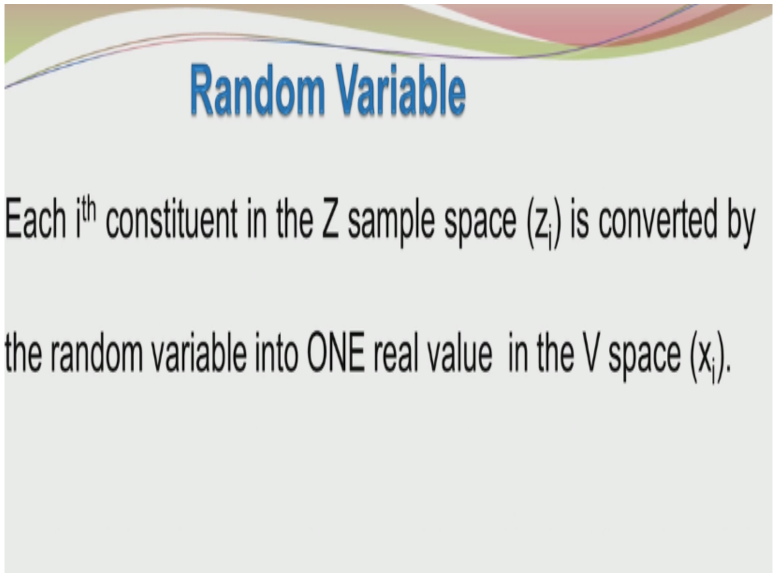
So you have a original sample space z and that is having entities zi, I running from let us say 1 to n z1, z2, so on to zn. These are the entities of the original sample space. Now the random variable function applies to each and every entity in the original sample space z and converts it into another value. So it assigns a single real value to each and every possible outcome in the original sample space z.

(Refer Slide Time: 16:06)



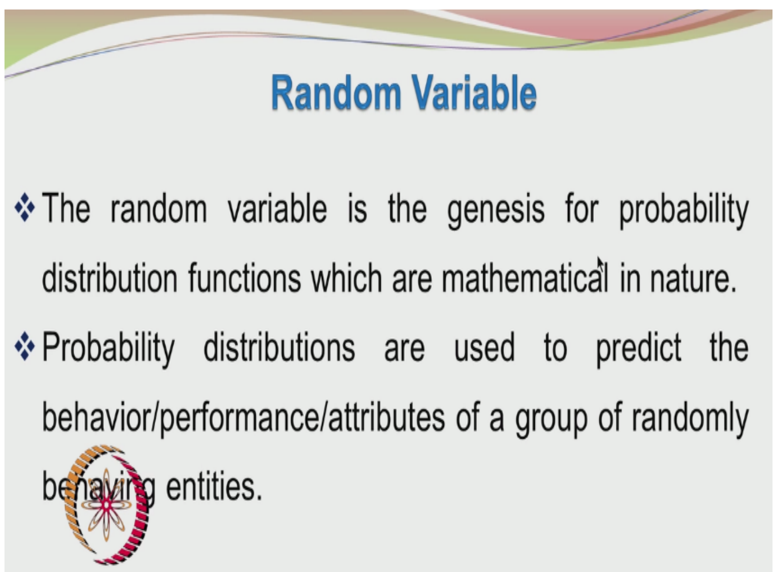
Why do we need to do it? The original sample space may have outcomes that are raw. It is like your raw data and they are tediously descriptive and inefficient to handle. So the random variable function converts these that is, all possible events in a sample space into real numbers held in the random variable space. Ogunnaike's book on Random Phenomena has a very nice discussion on these issues. Remember the most important thing is the random variable function converts these tedious and inefficient t handle outcomes into real numbers and they are held in the so called random variable space.

(Refer Slide Time: 17:08)



Each ith constituent in the Z sample space zi is converted by the random variable into one real value in the random variable space which is also called as the V space. So, the original sample space is referred to as Z and the random variable space is referred to as V.

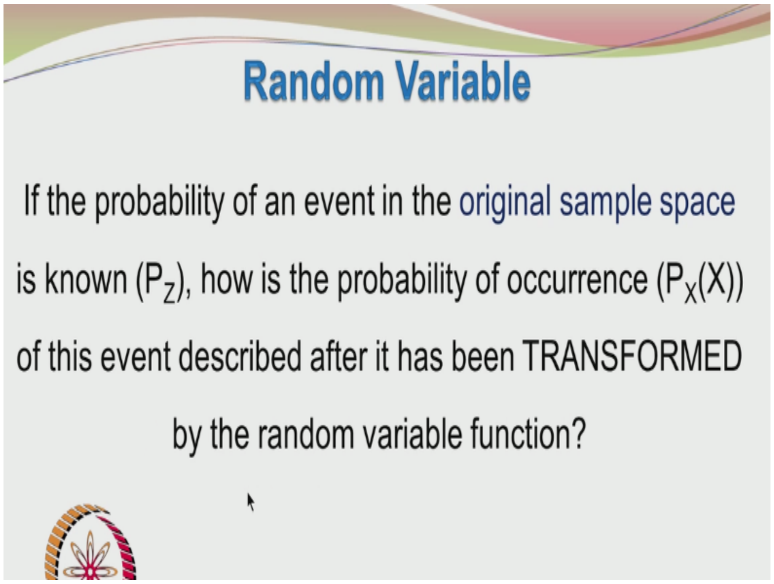
(Refer Slide Time: 17:37)



The random variable is the genesis for the probability distribution functions which are mathematical in nature. Why do we need the probability distributions? We cannot be doing experiments all the time to see the occurrence of even numbers or odd numbers, so once we have understood the random behaviour we have identified the random variable.

We can see the range of values it can take and we can also model the probability of the values that may arise from the experiment. We have a probability distribution and once we have a probability distribution it becomes easier for us to do further calculations. In a sense, we are having the collection of random variables and the behaviour or the performance or the attributes of this group of random variables may be predicted by the probability distribution.

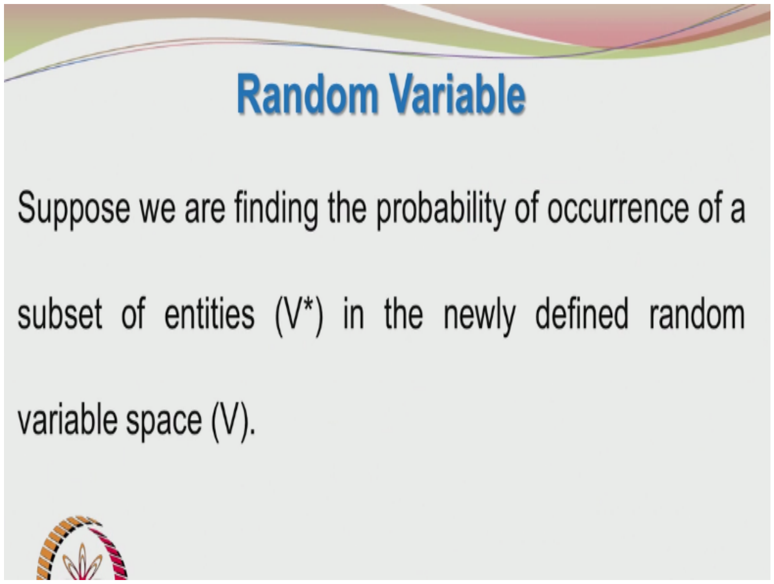
(Refer Slide Time: 18:54)



If the probability of an event in the original sample space is known and that is given by P of z how is the probability of occurrence of this event described after it has been transformed by the random variable function. (()) (19:20) we think look the event is uneven irrespectively of whether it is in the original sample space or it is in the random variable space the probability values should not really change.

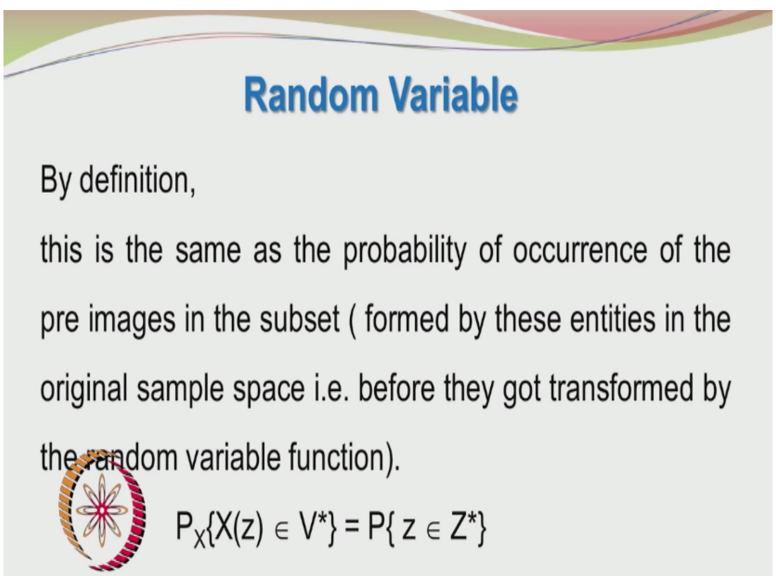
It is like you are having wine in a bottle and then you are transferring the wine from the bottle from which it is difficult to drink for example into wine in a glass or jar. So the taste of the wine is not going to change whether it was in the bottle or it was in the glass or in the jar. So we expect the probabilities to be same. Let us see what happens.

(Refer Slide Time: 20:19)



So we are interested in finding the probability of occurrence of a subset of entities V star in the newly defined random variable space V. So we are having a subset of entities in a freshly created random variable space we want to find the probability of occurrence of these entities.

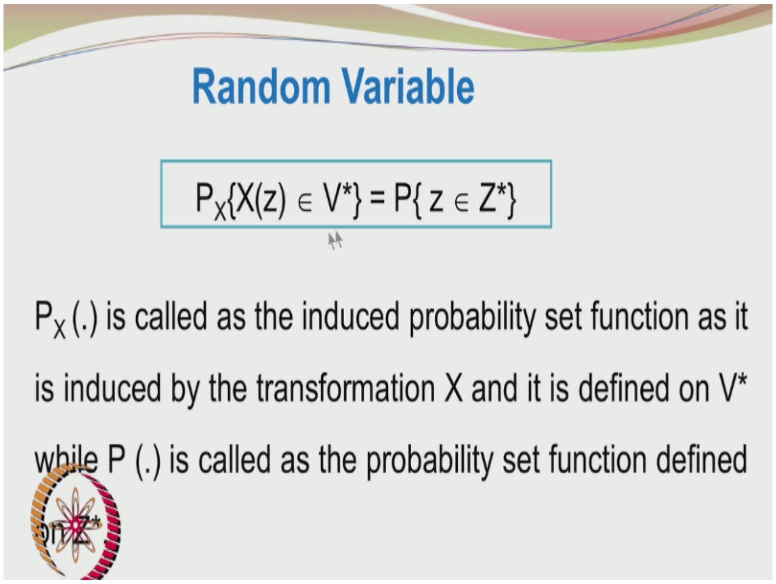
(Refer Slide Time: 20:46)



By definition, this is the same as the probability of occurrence of the preimages in the subset formed by these entities in the original space that is before they got transformed by the random variable function. Putting it mathematically, the probability of x of z belonging to V star which is the subset.

In the random variable space = the probability of the z which is belonging to the subset z star in the original sample space. Both the probabilities are the same. This represents the probability of the occurrence of the event in the transformed random variable space and this represents the probability of the event in the original sample space. Both the probabilities are the same.

(Refer Slide Time: 22:05)

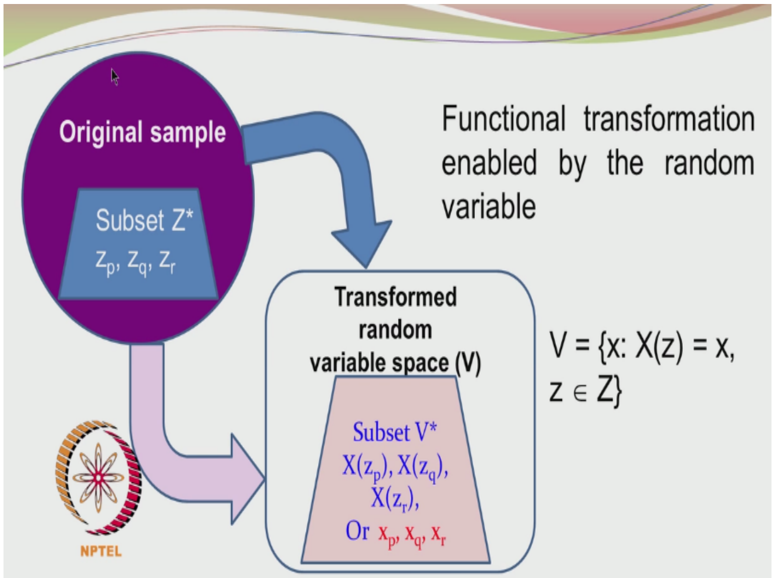


So we have used slightly different terminologies. This is the representation of the probability applied to the original sample space and this is the application of the probability to the transformed random variable space. So we denote p of x dot. This is the argument given here. This is called as the induced probability set function and it is induced by the transformation x and it is defined on V star.

We know that we have applied the random variable function to the subset z in the original sample space. Z are all elements belonging to the z star subset in the original sample space and after having applied the random variable transformation we have created a fresh subset V star in the random variable space and so this is termed as the induced probability.

When you want to calculate the probability of the events in the transformed random variable space the subset of events belonging to that transformed random variable space we call it as the induced probability set function and the other probability P without any subscript or superscript is called as the probability set function and it is defined on Z star which belongs to the original sample space.

(Refer Slide Time: 24:06)

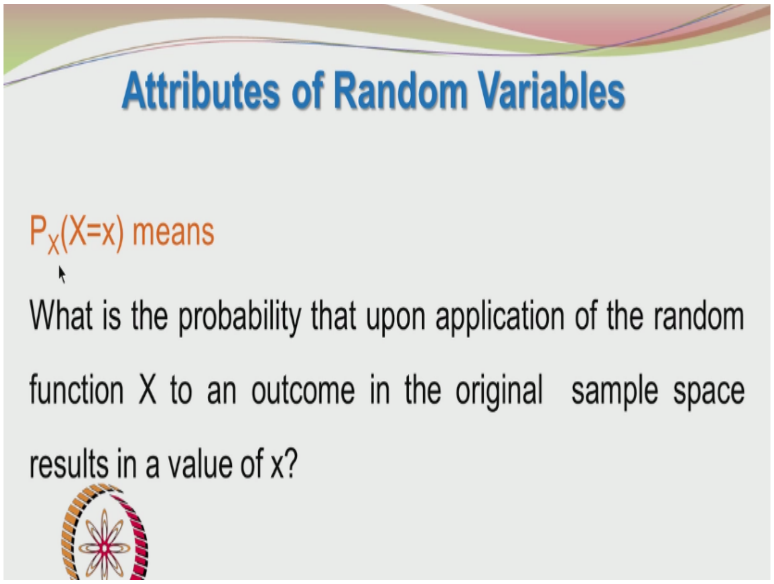


So whatever I have said now, I would like to represent in the form of diagram. This is based on Ogunnaike's Random Phenomena book. So, here we have the original sample space and there is a subset Z star in this original sample space containing elements Zp, Zq, and Zr. So these are three entities the original sample space and we have collected them into a subset Z star. Now look at this arrow. Using the random variable function, the original sample space has been transformed or converted into the transformed random variable space V.

So we have created a new random variable space V, by applying this transformation and so we have the subset also getting converted from subset z star the original sample space to subset V star in the random variable space. So we have the conversion x of zp. x of zq and x of zr and when we have these x of zp, zq, zr these are the random variable functions applied to zp, zq, and zr and they take the values small xp, small xq, small xr just as the subset z star had values zp, zq,

zr you apply the random variable function to each of these three entities and convert them into xp, xq, and xr. So this random variable V is defined as a set of all x not only these but also other values here which have been obtained by applying the random variable function x applied in all the z values in the original sample space. So we have created a new set by applying the random variable function. This is very interesting to look at all these transformations and even the random variable is termed as a function which converts the entities in the original sample space into random variables in a new space.

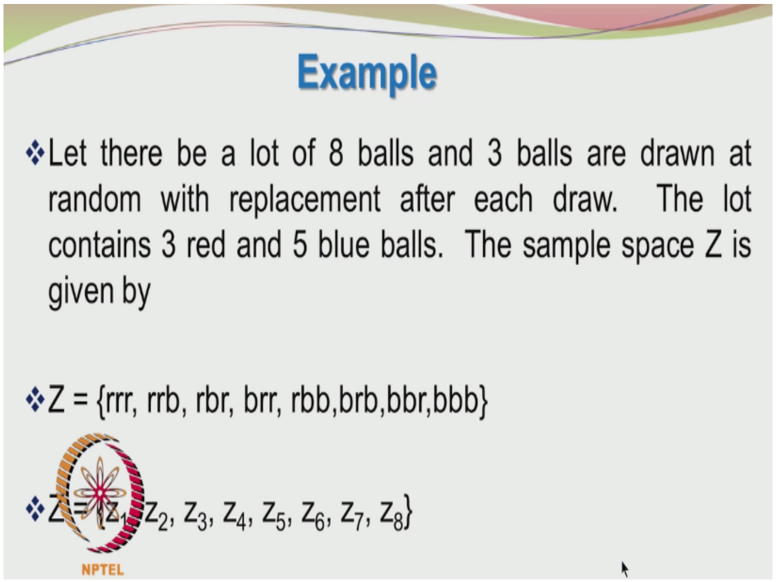
(Refer Slide Time: 26:56)



So we can now say that p of x where x is a random variable taking the value x we can describe it as what is the probability that upon application of the random function x to an outcome in the original sample space results in a value of small x. For some of you what I am explaining now may be interesting so you can read up more on this.

For some of you it may be a bit abstract highly mathematical, but it is not really very essential for you to understand it if you are finding it difficult. It is just to illustrate what is meant by a random variable I am going to this depth, but we are not going to really talk more about the induced probability function and the original probability function. It is just a preliminary introduction.

(Refer Slide Time: 28:01)



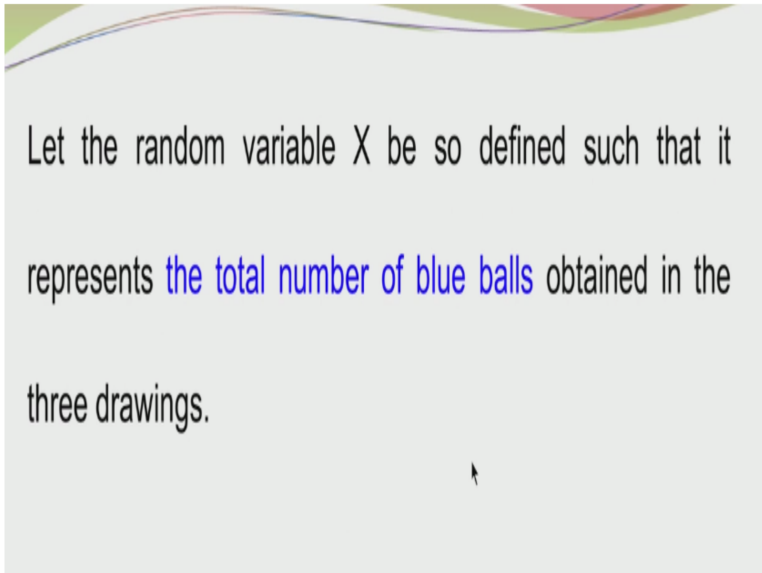
So now I will give an example to demonstrate the concept of random variable. It is a very popular example where you can find it referenced in many books is a standard problems and I am sure that you might have even come across it earlier in your academic career. So the example is there is a box or a lot of 8 balls and 3 balls are drawn at random with replacement after each draw. What happens is, you are having a box of 8 balls.

You take 1 ball, note the colour and then you put it back, then you take the second ball you take second time note the colour of the second ball then put it back. Similarly, you do it for the third ball note the colour and put it back. Important thing is after every inspection you are putting the ball back. Let us say that the lot contains 3 red balls and 5 blue balls and so the sample space is the list of all possible outcomes for the experiments you have conducted so even though you are conducting only 3 picking ups and replacing you can have several possibilities.

You may have all the balls to be red which are represented by r, r and r or you may have rrb, rbr, brr, rbb, brb, bbr, bbb and so on that is it. So you are having these as the possible outcomes. This is a tedious way of representing and it becomes even more tedious when you have for example more number of balls or you are having more number of experiments instead of doing for example 3 trials you may want to do 5 trials instead of 3 balls being drawn, you may want to draw 5 balls at random.

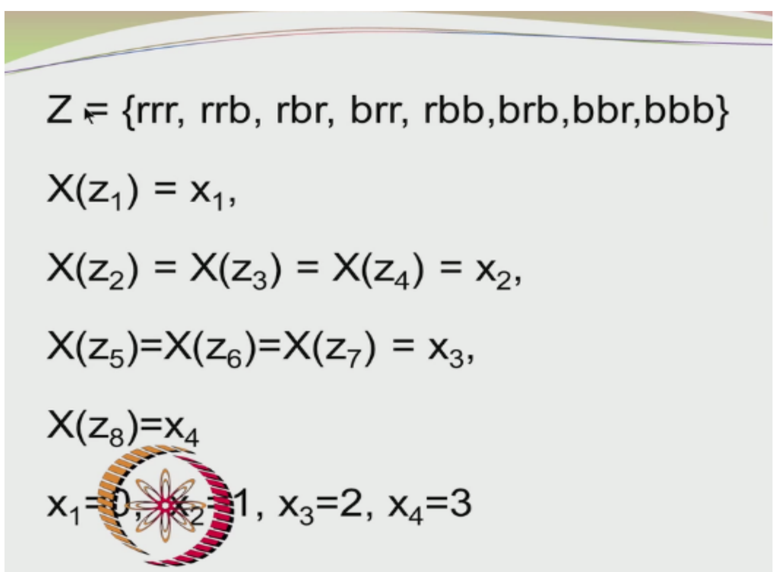
Then it becomes even more tedious to depict. So let us call each one of them as z1, z2, z3, so on to z8. So these are the entities in the original sample space.

(Refer Slide Time: 30:47)



We will represent it in the form of a random variable. Random variable is defined based on what we want. So it is here defined as the total number of blue balls that are obtained in the 3 drawings. So it represents the total number of blue balls which are obtained in the 3 drawings this is important.

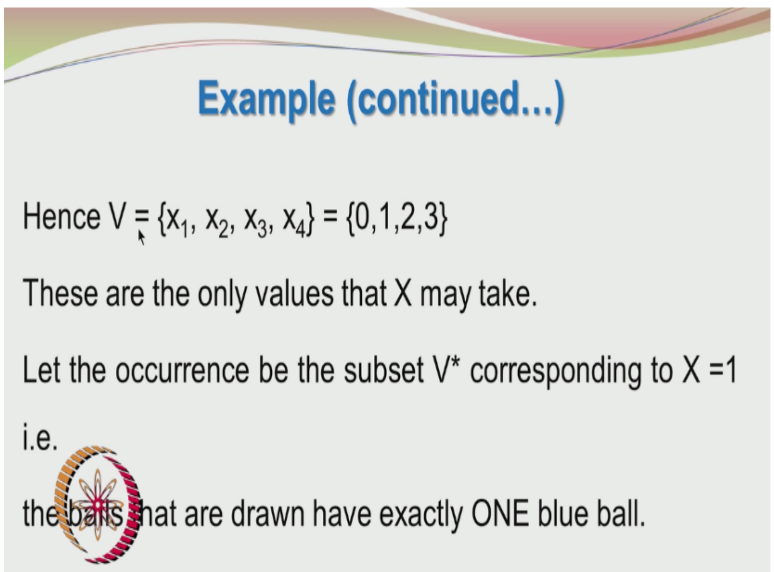
(Refer Slide Time: 31:15)



So you had the original sample space in terms of all these entities and we say apply the random variable x of z1 we apply the random variable function to this entity we get x1 and this is since our random variable is number of blue balls we apply the random variable function to these 3 and then we get x2. All these 3 represent occurrence of 1 blue ball and 2 red balls. They may not be in the same sequence, here gets rrb then rbr, then brr, but the fact in all these 3 outcomes.

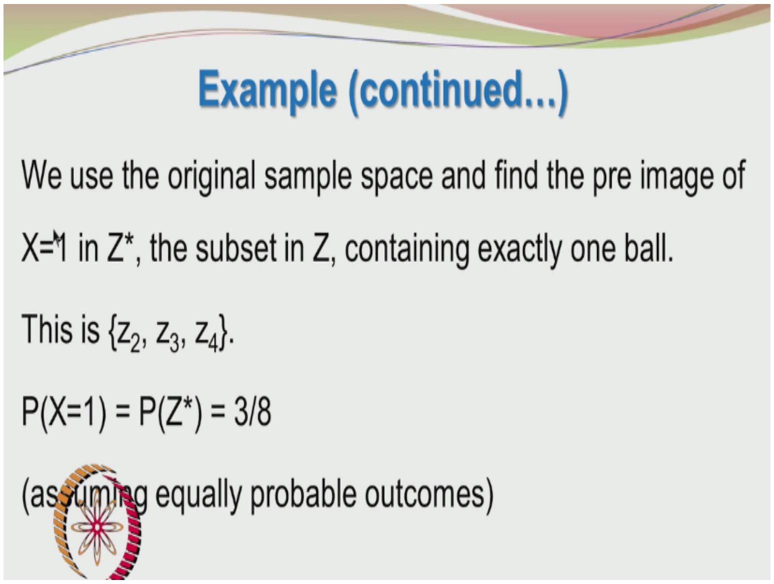
We are having only one blue ball so when you apply this we apply to z2, z3, z4 and then we call it as x2. We represent all of them all these events as x2 and similarly for 2 blue balls we apply the random variable function to z5, z6, and z7 and then we call it as x3 and x of z8 we call it as x4. So what are the values taken by x1, x2, x3, x4 it is no blue ball and there is only 1 blue ball in these 3 occurrences and there are 2 blue balls in these 3 occurrences so you have x3 taking a value of 2 and then finally you have x of z8 = x4. x4 is 3. Here we have all of them as blue balls.

(Refer Slide Time: 33:14)



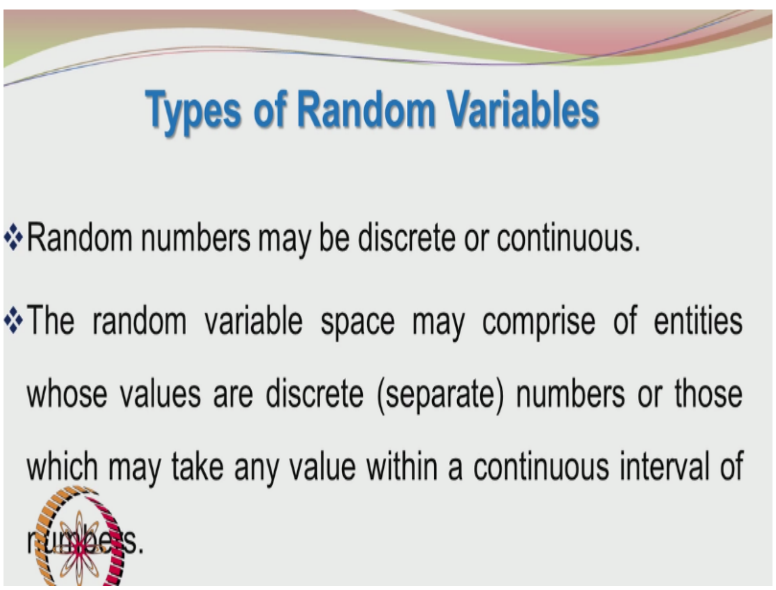
So the transformed variable space or the random variable space has 4 entities x1, x2, x3, and x4 and these may be taking the values 0, 1, 2, and 3 and these are the only values that x may take. Now let us define the occurrence be the subset V star corresponding to x = 1 that is the balls that are drawn have exactly only 1 blue ball so that is what we define as our subset V star.

(Refer Slide Time: 33:54)



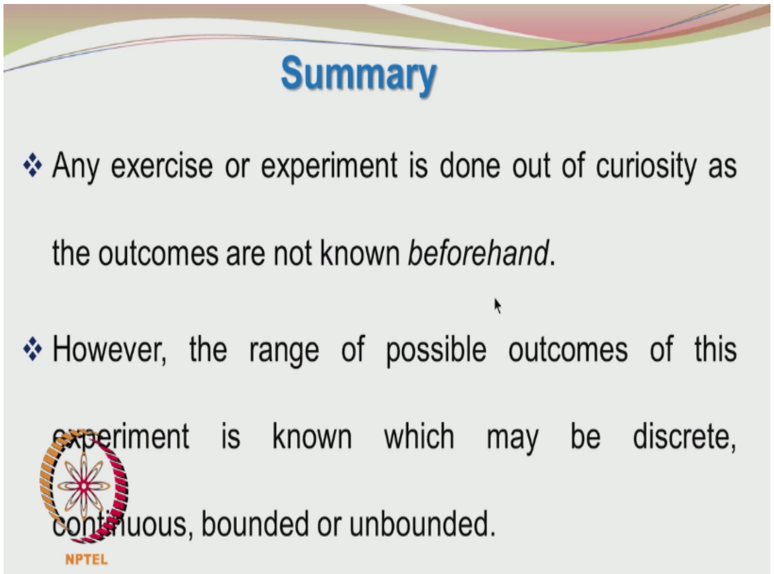
And in the original sample space the preimage of x = 1 and z star the subset in z containing exactly 1 ball is z2, z3, z4. Z2, z3, z4 all these had only 1 blue ball. So the probability of the random variable x taking the value 1 blue ball is = the probability of z star and the original sample space and that is = 3/8. We are assuming that all the balls have equal chances of getting selected.

(Refer Slide Time: 34:51)



So far we have seeing random numbers that have taken only 1 discrete value. They may also take continuous values and they may be lying within a certain interval the mole fractions of gases in a mixture range between 0 to 1 it is continuous set of values bounded between 0 and 1.

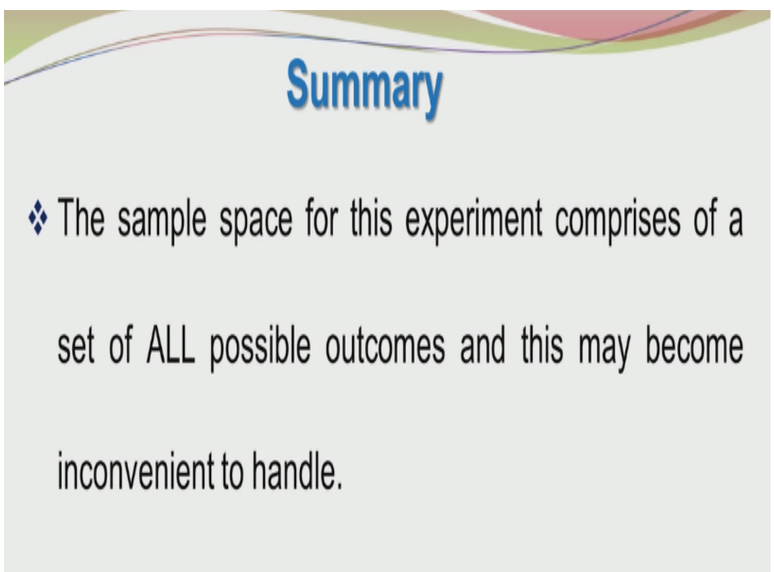
(Refer Slide time: 35:30)



Now coming to the summary of our lecture, whenever we do an experiment we do it out of curiosity. We will not be doing the experiment if we can predict the outcome. So since we are unsure about the outcome we carry out the experiments. We do not know what is going to happen in a game. So the game is played and then we get to know the outcome. So it is driven by curiosity and the outcomes are not known before hand and usually however the range of possible outcomes of this experiment is known which may be discrete continuous bounded or unbounded.

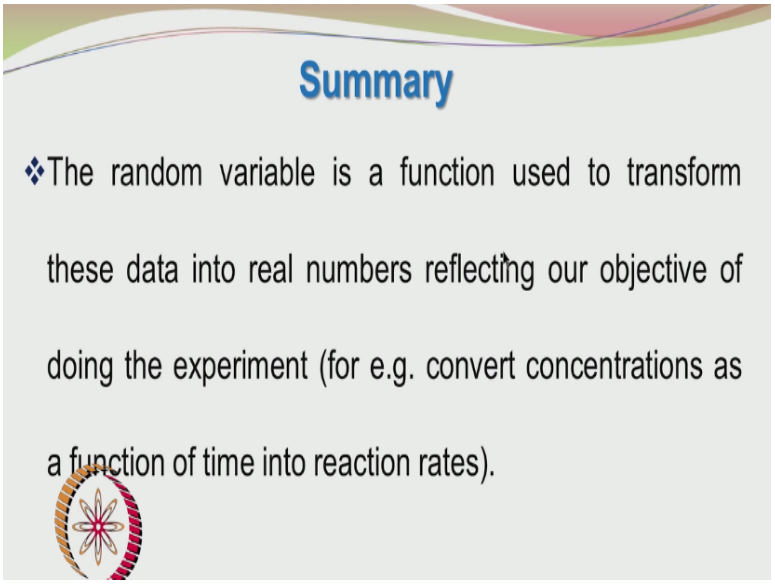
We can know the range of the possible outcomes.

(Refer Slide Time: 36:34)



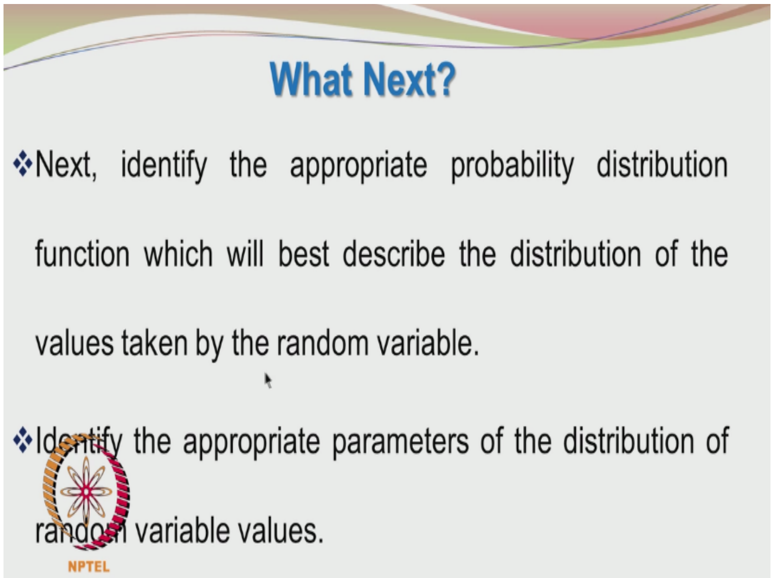
So we define the sample space for this experiment as a set of all possible outcomes and this may become inconvenient to handle to record and use.

(Refer Slide Time: 36:47)



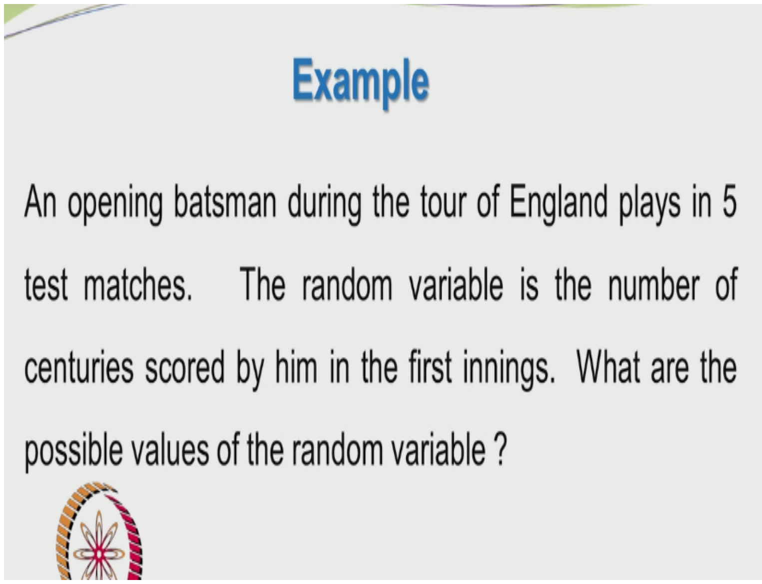
So we use a random variable function to convert these data into real numbers based on the objective for doing such an experiment. In the case we do reaction engineering experiments we want to convert concentrations as a function of time into reaction rates.

(Refer Slide Time: 37:10)



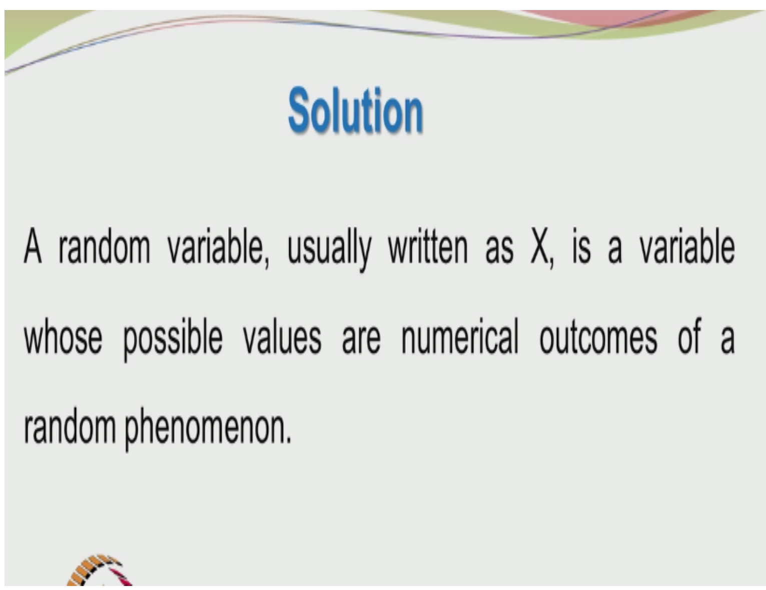
So what do we do next? We have been talking about random variables, the range of variables, values these variables may take and what we do next is to assign probability values to these possible outcomes and we are going to study the characteristic features of such probability distributions.

(Refer Slide Time: 37:45)



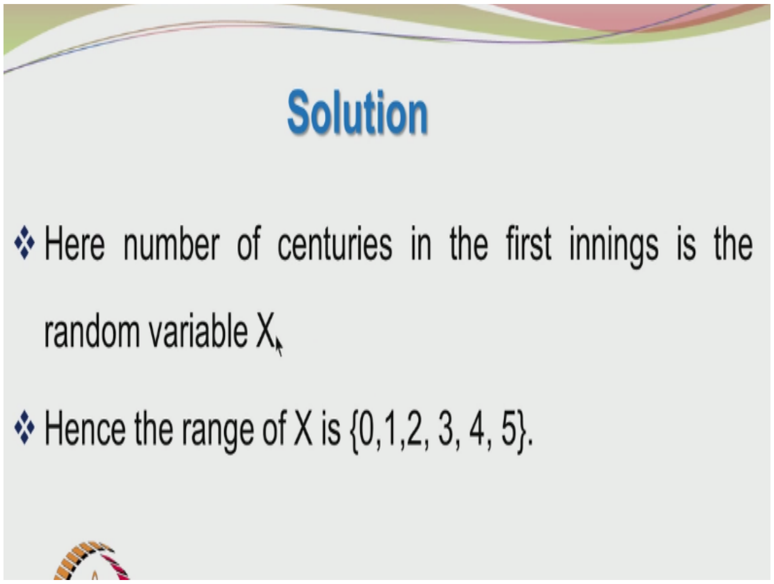
So let us look at few examples to illustrate the concept of random variables. So we will take an example of an opening batsman touring England. He plays in 5 test matches. We define the random variable as the number of centuries scored by him in the first innings so what are the possible values this random variable may take.

(Refer Slide Time: 38:14)



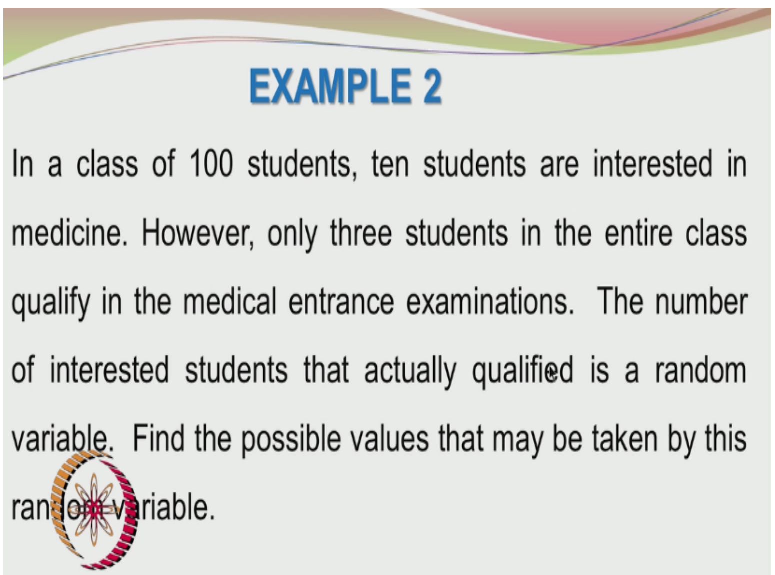
Random variable is represented as X and is less the possible values which are numerical outcome of a random phenomenon. So the random phenomenon here is scoring essentially and how many centuries the batsman scores during the 5 tests matches are the numerical outcomes.

(Refer Slide Time: 38:46)



So here the number of centuries in the first innings is the random variable x and so the range of x is 0, 1, 2, 3, 4, and 5. So it is between 0 to 5. Let us go to the next example.

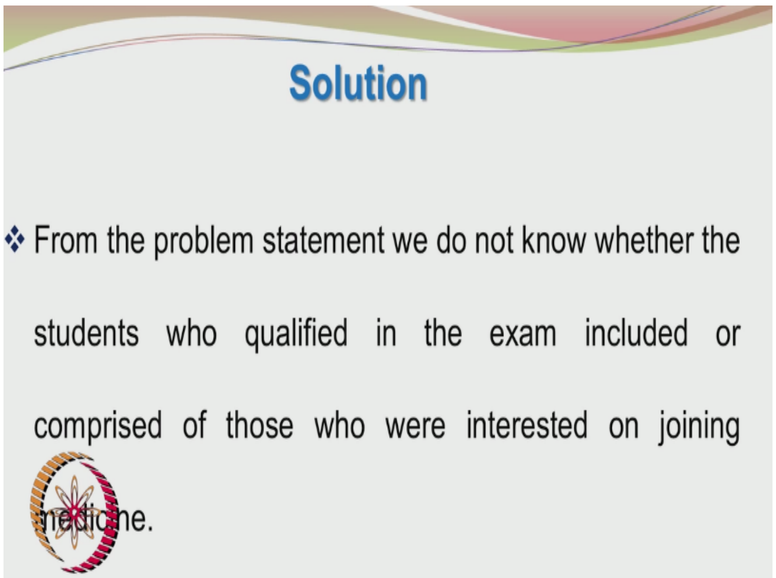
(Refer Slide Time: 39:09)



Here we have a class of 100 students and 10 students are interested to take medicine and only 3 students in the entire class qualify in the medical entrance examinations. So out of 100 students, 10 students have expressed interest to take or do medicine and 3 students in the entire class qualify in the medical entrance examination.

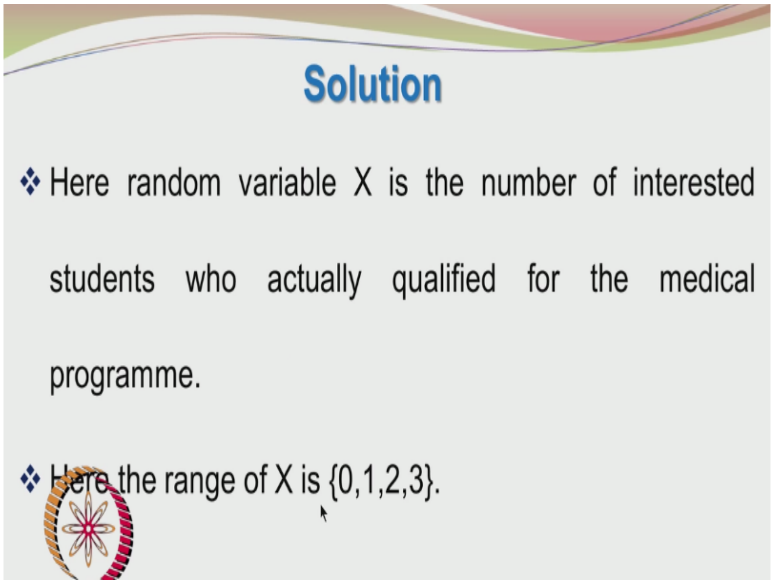
A number of interested students that actually qualified is a random variable. Find the possible values that may be taken by this random variable. So if we read the question again we see that the number of students who had expressed interest to take medicine and out of them how many have actually qualified, this I the random variable definition.

(Refer Slide Time: 40:13)



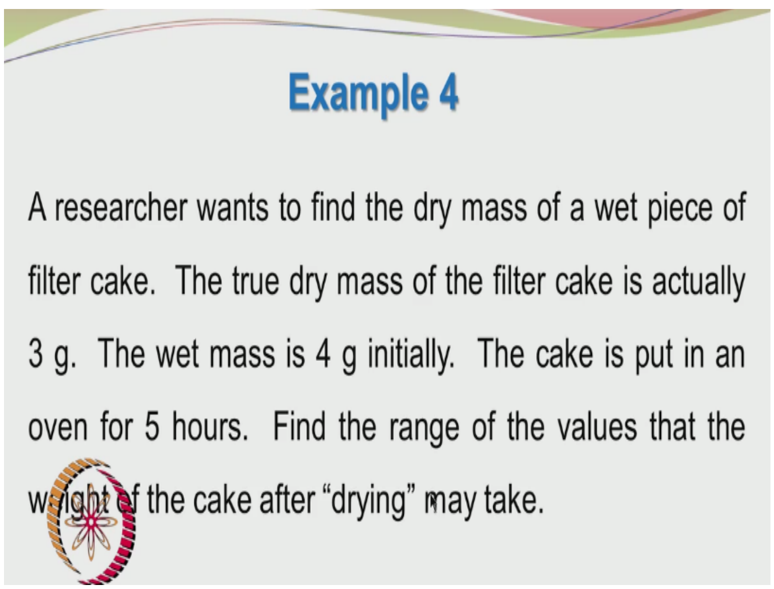
So as the slide says we do not know whether the students who qualified in the exam included or comprised of those who are interested on joining the medicine.

(Refer Slide Time: 40:26)



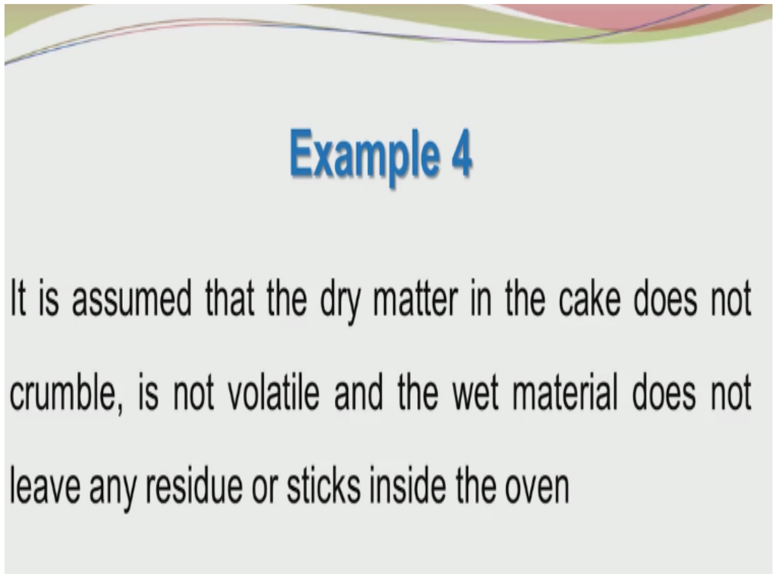
So the values that may be taken by x is 0, 1, 2, and 3. What this means is if the random value variable value 0 it means none of the 10 students who had expressed interest to take up medicine had actually qualified in the exam. If this is 3, 3 out of those 10 students had qualified for the medical program.

(Refer Slide Time: 41:05)



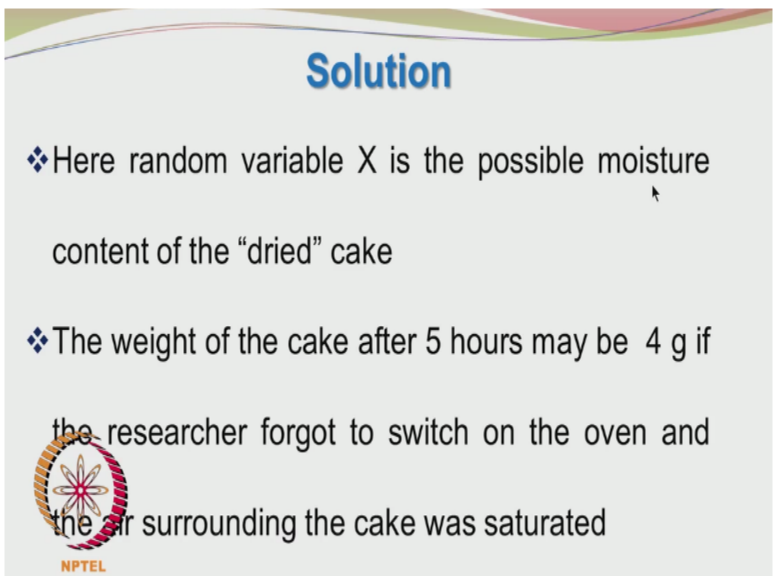
The next example, the researcher wants to find the dry mass of a wet piece of filter cake. The true dry mass of the filter cake s actually 3 g. The wet mass is 4 g initially. The cake is put in an oven for 5 hours. Find the range of the values that the weight of the cake after drying may take?

(Refer Slide Time: 41:33)



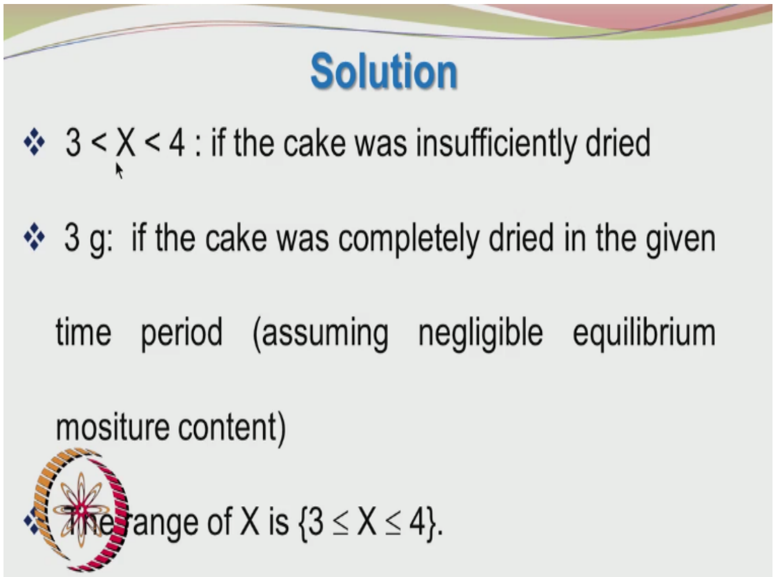
So we make the assumption that the dry matter in the cake does not crumble, it is not volatile and the wet material does not leave any residue or sticks inside the oven

(Refer Slide Time: 41:48)



Here random variable X is the possible moisture content of the dried cake. The weight of the cake after 5 hours may be 4 g if the researcher forgot to switch on the oven or the air surrounding the cake was saturated so no evaporation occurred. So it may be still 4 g. There is no weight loss in the cake.

(Refer Slide Time: 42:15)



The cake may have been only insufficiently dried so the weight of the cake may take values between 3 and 4 or the cake might have been completely dried to reach the bone dry mass of 3 g. We are assuming negligible equilibrium moisture content so the cake was completely dried and became 3 g. So the random variable x may take values between 3 and 4, both 3 and 4 or included. This concludes our discussion on random variables.