Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology – Kharagpur

Lecture – 03
Random Variables

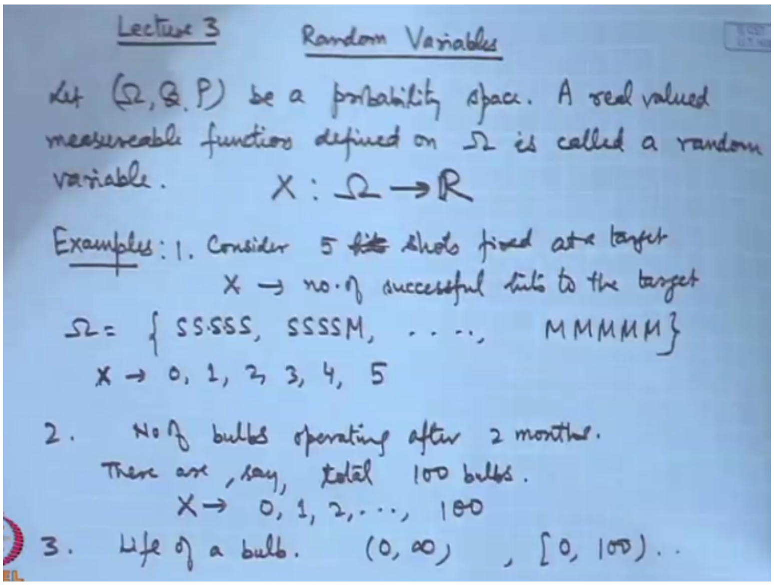
In the previous lecture, we have discussed various laws which govern us to calculate the probabilities of various events one thing that we noticed that, when we are discussing the raw sample space many times, it is quite difficult, or you can say complicated to look at all, the possibilities and then look at the manifestations of various events. It may happen that it may be helpful to look at only the numerical values of certain variable.

For example, let us consider the game of a say a basketball match. Now at the end of the match, we will be interested in the total number of the baskets which we were put successfully by both the team and that would give the winner in the match similarly, in a game of tennis, in a game of badminton, and so on where the outcome may be simply a number.

For example, in a game of badminton the final scores of the 3 games or 5 games or in a game of tennis, the final scores of best of 3 games or best of 5 games etc. So, now the entire duration of the match is not important, which point was won by which player that is not important, okay. That means ultimately we associating certain numerical values with the phenomenon or you can say outcome.

Now this association if we formalize for all values of the sample space we call it random variable. So, let us formally define what is a random variable?

(Refer Slide Time: 02:03)



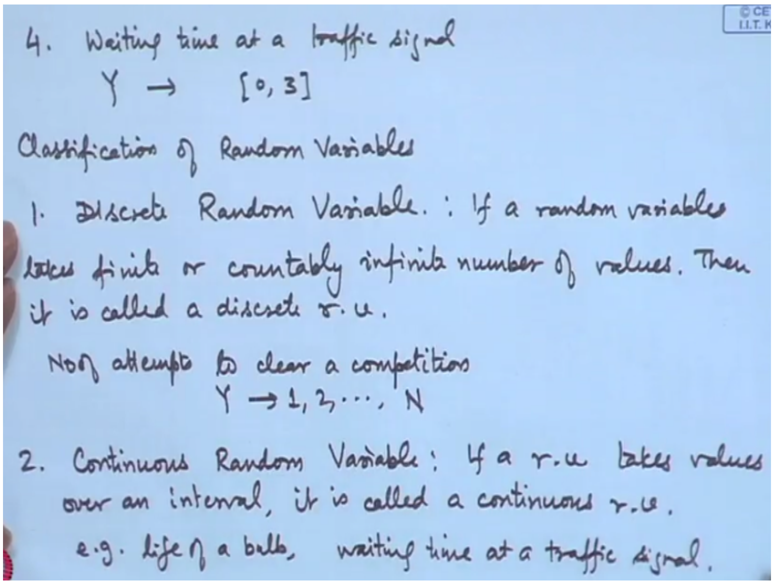
So let omega B, P be a probability space. A real value measurable function defined on omega this is called a random variable. That means so now usual notation for the function in mathematics is FG etc however here it is random variable so we call it as X, Y, Z etc. So we use a notation for example, X is a function from omega into the set of real numbers at the same time we put measurability condition.

By measurability we mean that if we consider a sigma field of subsets of R then the inverse image of any of the set should be an event here that means it should be a set here. Now, typical example let us consider so consider 5 shots fired. Now out of this x could be the number of successful hits to the target. In that case, see when the 5 shots are fired at a target then the sample space may have the possibilities like all are successful so you may have 4 successful and one missed and so on all are missed.

These are the possibilities. However, X can take only values 0, 1, 2, 3, 4, and 5 and so here you can see that looking at this one is quite easy and convenient. Let us take some more examples. Suppose we are considering the number of bulbs operating after say 2 months. So now this could be for example, there may be there are say total 100 bulbs. Then how of them will be operating after 2 months.

So let us say x is the number here then x can take value 0, 1, 2 up to 100 and we may look at the probability distribution of this. We may look at the life of a bulb. So in this case it could be some number say from 0 to infinity or it could be 0 to 100 and so on. It can be depends upon the wave we try to use our sample space.

(Refer Slide Time: 05:56)



We can consider say waiting time at a traffic signal. Suppose this is y then waiting time can be anything from say 0 to 3 minutes. Suppose if the timer is set up to 3 minutes then it could be 0 to 3 minutes. The possible values can be any value between 0 to 3 etc. So here you can see that in each of these examples we are having a numerical value from the sample space that means each element of the sample space is allotted numerical value.

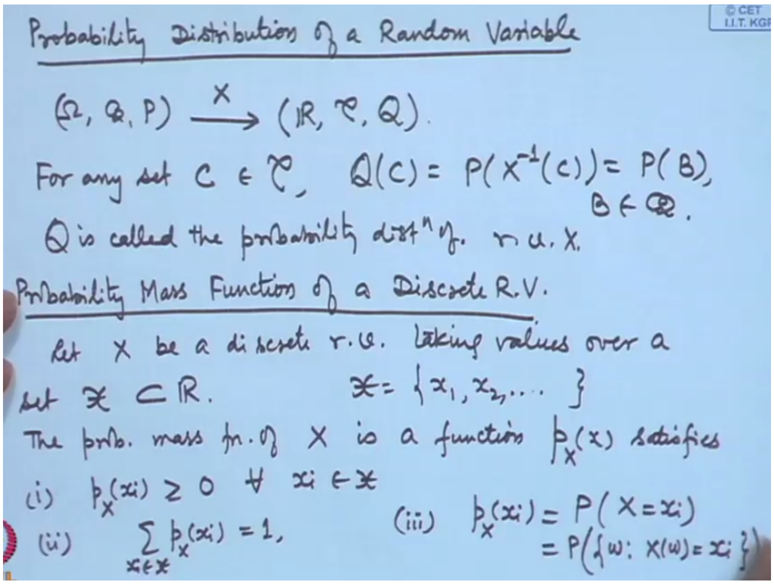
So this is called random variable. Now based on the type of the values for example here I am assigning the values 0, 1, 2, 3, 4, 5. Here I am assigning 0, 1, 2, up to 100 and so on. Here I am assigning the values on an interval so depending upon that we consider the classification of random variables. So the classification of random variables is one is discrete random variable.

So if a random variable takes finite or countably infinite number of values then it is called a discrete random variable. For example, if you look at this random variable this is a discrete random variable that is the number of successful leads to the target. Suppose I consider number of attempts to clear a competition. So a person may be allowed finite number of attempts to clear the competition.

So the values that the random variable may take 1, 2 suppose up to N number of trials are allowed so you can have 1 to N as the number of values of Y. Then continuous random variable if a random variable takes values over an interval it is called a continuous random variable. So for example waiting time at a traffic signal I am putting 0 to 3 interval, or life of a bulb 0 to infinity etc.

So this is the example of continuous random variable for example, life of a bulb, waiting time at a traffic signal. There are also random variables which are partly discrete and partly continuous, but that distinction will become clear when I give the probability distribution of a random variable. So, let us look at the probability distribution now.

(Refer Slide Time: 09:58)



How do we define the probability distribution of random variable? First of all let us look at the origin of this one. See whenever we have a random experiment, from the random experiment we have a sample space and certain assignment of probabilities can be made depending upon the type of experiment to the various outcomes or to the various events which are there.

Now when we are allocating the values, real numbers to those outcomes then the corresponding probabilities are also transferred. Now in a case of discrete distribution this assignment is quite natural, but in the case of continuous distributions since we are dealing with the intervals that assignment does not result in a direct transformation. However, we can give a formal definition in the following fashion.

So we can consider, so we have omega B, P as the probability space and when we define a random variable X then omega goes to R. This sigma field of events it is also transformed to another sigma field which is a sigma field of subsets of R and in place of P let us say Q. So this is an assignment let us define. So for any set C we define QC = probability of X inverse C = probability of B where B belongs to B. So here Q is called the probability distribution of random variable X.

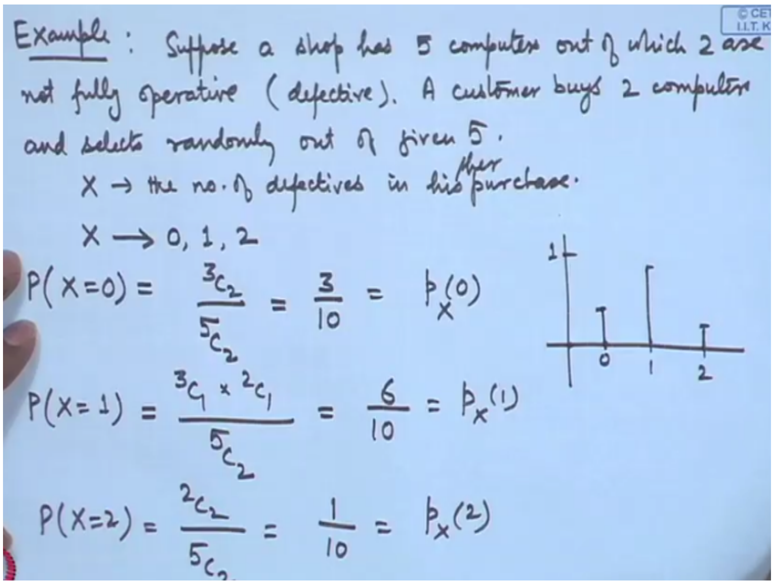
Now let us see that how it is done in the real situation when we dealing with the discrete random variable when we are dealing with the continuous random variable then we will have different ways of this assignment. So let us consider the discrete case. Here this probability distribution is given the name probability mass function of a discrete random variable. So then it is a discrete random variable.

The random variable takes values over a set which is either finite or it is countably infinite. We can name it by some script X. Let x be a discrete random variable taking values over a set say X which is of course a subset of R. So here x can be since it is finite or countable we can actually arrange the values in this sequence x1, x2, and so on. Now the probability mass function of X is a function let us call it Px which satisfies the following 3 conditions.

First thing is that Px(xi) is always > or = 0. Second is that the sum of all the possible values of the function over x = 1 and thirdly, this Px(xi) is actually the probability that the random variable X takes values xi. Now what is the meaning of this kind of statement let us explain it further. I wrote here the assignment of value here now this set B is the subset of omega okay the sample space.

So when we writing like this actually it means probability of all those sample points for which x omega = xi. So this is actually the full description, but for gravity we will be actually writing x = xi because probability statement is basically valid for a set on the which is a subset of the sample space so omega the set of all those sample points for which x omega = xi. So in short, we will be writing probability of x = xi, but when the meaning is clear it does not matter whether we write this statement or this statement. Let us explain through an example here.

(Refer Slide Time: 15:28)



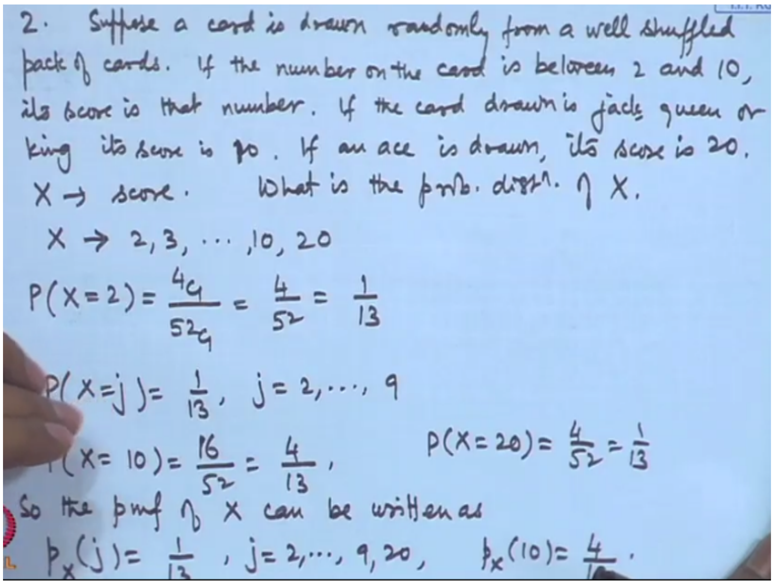
Suppose a shop has 5 say computers out of which 2 are not fully you can say they are not working up to the full capacity not fully operative okay. You may call that they are having some defect okay. A customer buys 2 computers and selects randomly out of given 5. Because that they are defective that is not known to the owner also. Now let us consider X is the number of defectives in his or her purchase.

So now what are the possible values of X? Since there are 5 out of which 2 are defective and this person buys 2 he may buy both good, he may buy one bad, one good or he may buy both bad what are various possibilities here. Let us look at what is the probability of $x = 0$ that is equal to that means he has purchased 3 c2 that means he has purchased both from the good ones the total number of ways of selecting 2 computers out of 5 is 5 c2 which number is equal to 3 by 10 here.

This is actually Px(0) similarly if we consider probability of $X = 1$ that means he has purchased one good and one bad so the probability here is becoming 6/10 that is px(1) and what is the probability of $X = 2$ here so that will be equal to he has purchased both bad that is 2 C 2/5 C 2 = 1/10 = Px(2). so this assignment of values that is Px(0), Px(1), Px(2) this is called the probability mass function of this random variable x here.

If we want, we can have a basic understanding by considering the plot of this. Suppose this is on so this is 0, 1, 2. So at 0 you have 3/10, at 1 you have 6/10 and then you have 1/10. So this can be depicted by bar diagram actually. Let us consider one more example of the discrete distribution.

(Refer Slide Time: 19:04)



Suppose a card is drawn randomly from a well shuffled pack of cards. Now we define the random variable X as follows if the number on the card is between 2 and 10, its score is that number. If the card drawn is Jack, queen, or king that is a picture card it is score is 10 and if an Ace is drawn its score is 20. So let us call x as the score okay. I want the probability distribution of x.

So what are the possible values of X here? X can take values if it is between 2 to 10 it is score is that number that is 2, 3 up to 10. If it is Jack, queen, or king then also it is 10 and if an ace is drawn its score is 20. So the possible values are 2, 3, 4, up to 10 and then 20. Now if you want to find out the probability distribution here what is the probability of say X = 2. Now that is simply = because there are 4 cards which are having the value 2.

And we are drawing one card out of 52 cards so it is becoming simple 4C1/52C1 = 4/52 = 1/13. But this argument will be valid for 3, 4, up to 9. So actually we can say probability x = J = 1/13 for J = 2 up to 9. Now the case of 10 can be clubbed with Jack, queen, and king because that score = 10 Now there are 16 cards which are 10, Jack, queen, or king, so the probability that x = 10 that will become equal to 16/52 = 4/13.

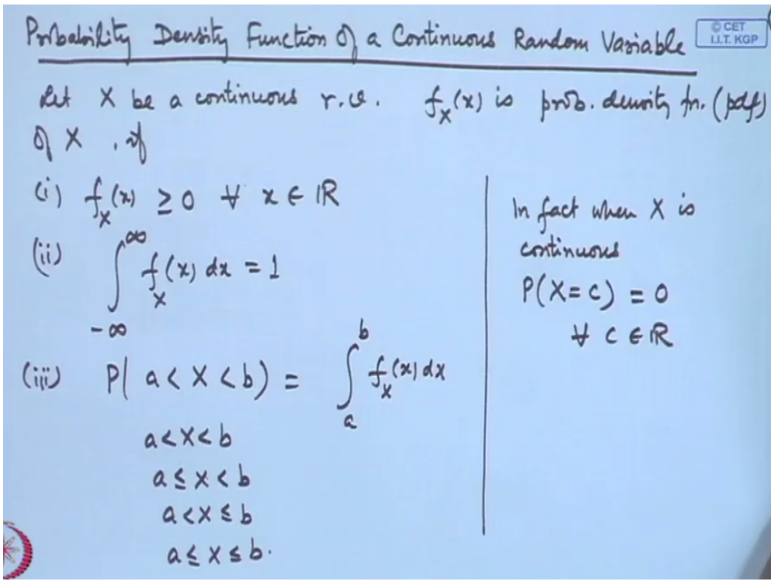
And what is the probability of x = 20 because that is an ace which is again = 4/52 = 1/13. So this number I can also combine here so ultimately we can write the distribution in the following form. We can write so the probability mass function of x can be written as Px(j) =

1/13 for j = 2 up to 9 and 20 and Px(10) = 4/13. So this is the complete probability distribution of the score which is defined in this fashion in this particular problem.

Now if we have continuous random variable it is obvious that one may not be able to get the probability distribution in this fashion. Because here what we are doing is we are making a simple assignment and since the elements of the sample space have a probability assignment that is directly transformed to the probability distribution of discrete random variable in the case of continuous random variable.

Since there are the values are taking over an interval now this type of facility is not there and in that case the probability distribution is obtained either by the rule which is governing the process or one may look at the probability histogram of the data that means frequency distribution and then we approximated by a frequency curve.

(Refer Slide Time: 24:13)



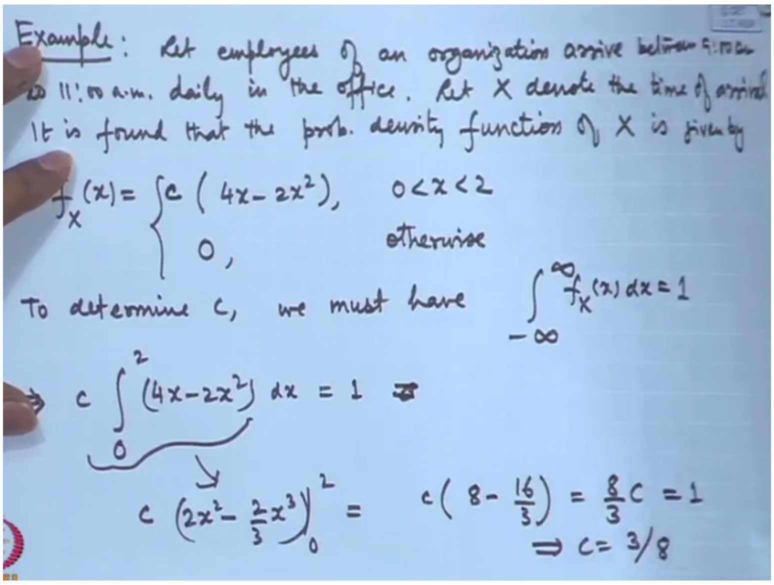
So let us consider probability distribution of a continuous random variable. In the case of continuous random variable, we call it a density function. So, probability density function of a continuous random variable. So let X be a continuous random variable and when it is a continuous random variable it is taking values over an interval. Now that interval can be having a closed and bounded interval or it can be infinite from both the sides so in general we can consider the set of values from - infinity to infinity.

Now over the portion where the probability is not there the density function will be defined to be 0. So we can define in general fx(x) is probability density function we call in short pdf of

X. If it satisfies that fx(x) is always > or = to 0 the integral over the real line = 1 and thirdly if we consider any interval probability of say a < x < b that is given by from a to b. Now in this case a < x < b we may also consider a < or = x < b we may consider a < x < or = b or a < or = x, or = b. all of them are considered to be equivalent.

In fact, when x is continuous probability of random variable taking a fixed value is always 0 for all c. In fact, this is not the density function. In the discrete case this is the probability mass function, but in the continuous case, this probability is always 0 actually you have a density at C which may be 0 to nonzero. So let me explain it through some example here.

(Refer Slide Time: 27:11)

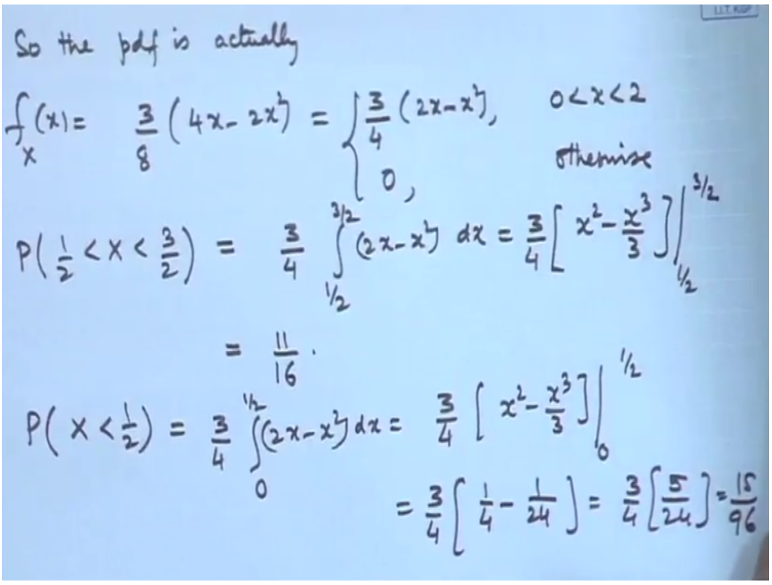


Let implies of an organization arrive between 9:00 a.m. to 11:00 a.m. daily in the office. Let x denote the time of arrival. Now here you see that we are having a 2-hour period and it is found that the probability density function of x is given by so considering the starting time as 0 the time can go up to 2 so the density function is given by fx(x) = C (4x - 2x square) 0 < x < 2 and = 0 otherwise. By otherwise we mean that when x is outside this interval.

Now that is natural because we are considering the arrival timings between 9 to 11 only so it is a 2-hour period so the probability density function is having a non negative value in this region only. Outside this it has to be 0. Now here there is an unknown constant c written here. So first thing is that let us determine this C. Now this function has to be positive which we can obviously see it is actually positive between 0 to 2.

Now secondly we observe that the integral of this function to determine C we should have the integral of fx(x) dx = 1 which is equivalent to saying c times integral 0 to 2 4x - 2 x square dx = 1. Now if you look at this integral so this becomes simply 2 x square - 2/3 x cube from 0 to 2 * c = 8 - 16/3 = 8/3 C now if this = 1 this means that c = 3/8. So actually the probability density function is now.

(Refer Slide Time: 30:28)



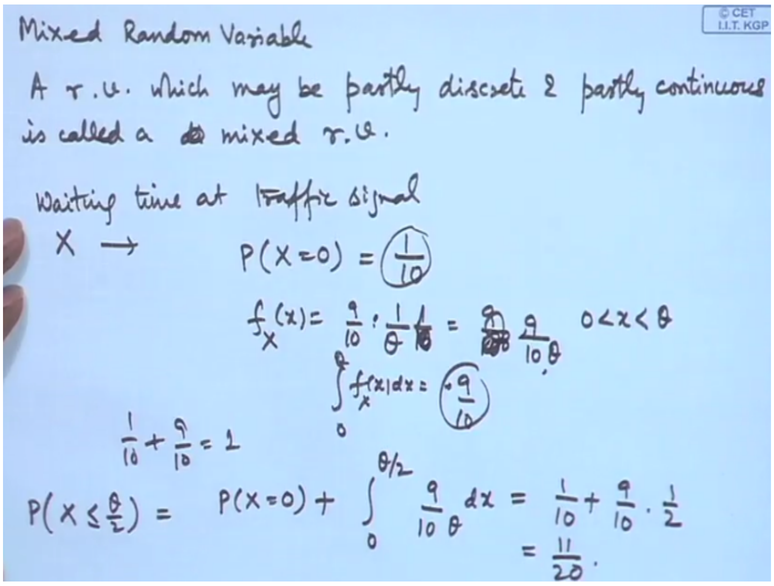
So the probability density function is actually fx(x) = 3/8 * 4x - 2x square or we can take 2 common here then it becomes 3/4 * 2x - x square for 0 < x < 2 and = 0 otherwise. Now suppose we are interested in certain probability for example what is the probability say that half < x < 3/2 that means we want to know that imply arrives between 9:30 to 10:30 time.

So then this is the thing but the integral of 3/4 2x - x square dx from half to 3/2 which of course can be evaluated easily which is x square - x cube/3 from half to 3/2 so we can substitute these values here and we can get the answer as = 11/16. For example, we want to find out what is the probability that x is < half that means what is the probability that an imply arrives between 9:00 to 9:30.

Then this will be equal to 3/4 0 to half 2x - x square = 3/4 and once again it is x square - x cube/3 0 to half so once again if you want we can actually evaluate this 3/4*1/4 - 1/24 = 5/ 24 = 15/96. You can actually see that the probability of arriving within first half of hour is much less there is more probability concentrated from half to 3/2 more implies arrive between 09:30 to 10:30. As I mentioned that there may be random variable which may take values over an interval and it may also take at a finite number of points in between.

Now what is the meaning of that? Now here you can see that when you are having a continuous random variable then the probability of a point is 0 whereas in a discrete case we have observed that the probability assignment is for the points only. Probability of x = xi = px(xi). Now we may have a random variable which may take positive probability at points and then for points it may be 0 and probabilities over intervals are allocated. This is called mixed random variable so let us look at this.

(Refer Slide Time: 33:46)



A mixed random variable. So a random variable which may be partly discrete and partly continuous is called a mixed random variable. Let us look at some physical example. We have considered say waiting time at a traffic signal. Now the vehicle is going and then there is a traffic signal so we are looking at the probability distribution of the waiting time. Now one may say that it is an interval depending upon the setting of the signal timing like it may be most of the signals are set up to 2 minutes, 3 minutes, or one-minute timing.

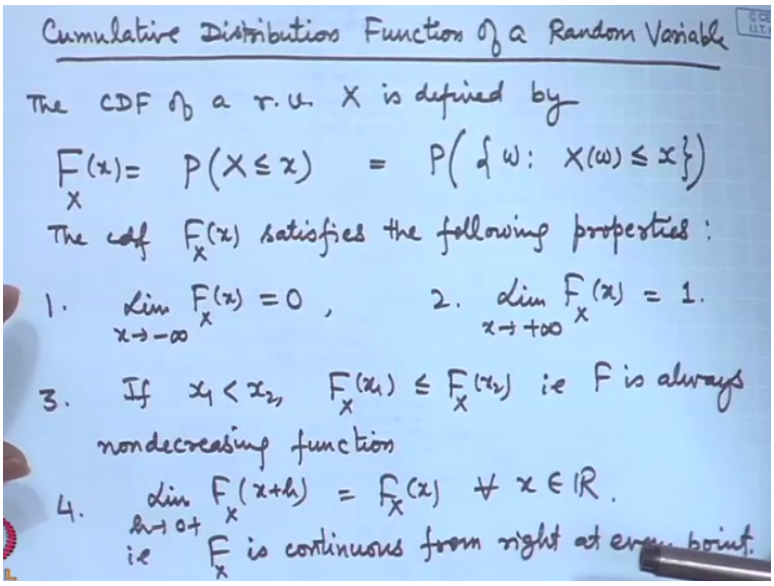
Then you may set the time is from 0 to 3 minutes and it is an interval but it may happen that when the vehicle is passing or when it is encountering the signal there is a green light and then the waiting time is actually 0. Now it may happen say one third of the time or one 4th of the time or one tenth of the time you actually encounter a green signal. So there is a positive probability of x = 0.

So you say for example waiting time at traffic signal so let us call it x so we may have probability x = 0. Let us say suppose one tenth of the time we observe this and thereafter we may have a density and that density may be having say an interval you have 9/10 * 1/10 = 9/100 and you are having 0 < x < sorry it could be some intervals say 0 to say theta so we may have 9/10 theta so that means that this is the total length of the interval in which you may have to wait.

So here you can see if I integrate this density fx dx from 0 to theta I get 9/10 now 9/10 + 1/10 = 1. So, the total probability assignment is completed but here you can see that in this range you have density function whereas at 0 you have a discrete point. So for example, a question may be asked what is the probability that the waiting time is < or = say theta/2. In that case this is equal to probability of x = 0 + the interval.

The integral of the density over the interval 0 to theta/2 so 9/10 theta dx so this = 1/10 + 9/10 then you have half here. So that is equal to 11/20. So this is an example of a mixed random variable. Now from this example it is clear that sometimes when you have a mixed random variable it may not be very convenient to consider the probability density function or the probability mass function. In that case, we have a more general function which is called the cumulative distribution function of a random variable.

(Refer Slide Time: 37:52)



Cumulative distribution function of a random variable. So the cumulative distribution function we can write in short CDF. The CDF of a random variable x is defined by Fx = probability of x < or = x. As I mentioned actually this is the probability of the set of all those points for which x omega is < or = x, but in short we can write as a probability of x < or = x.

So the name cumulative distribution has come because you are considering the probability of all the points which are coming < or = x.

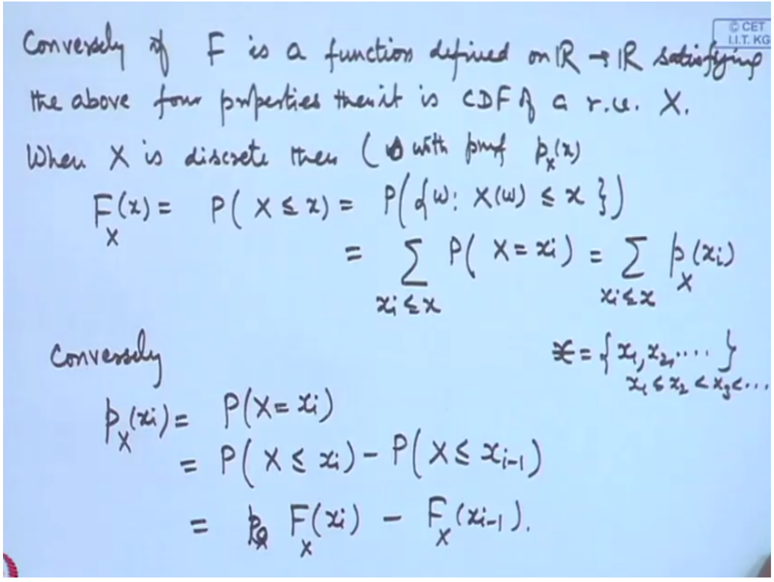
So you are adding for example if you say x = 0 then you are considering all the points which are up to 0 you are considering the total probability assignment up to 0. Then if I say x = 1 then further you are adding the probabilities of all the point which are assigned values between 0 to 1 and the previous values are already there that is why it is becoming cumulative distribution function.

We will see that this function is quite useful and also it is having certain characterizing properties. Now certain things you can observe easily. For example, if I say x tends to + infinity. If I say x tends to + infinity then naturally all the points will be incorporated here. So this will be actually become the probability of the full sample space and therefore this will be equal to one.

If I say x tens to - infinity then naturally this is showing that the now values can be assigned here and therefore the probability of the impossible event that will be equal to 0. So similar properties we can write here the cdf fx satisfies the following limit of fx as x tends to - infinity is 0. Limit of fx as x tends to + infinity = 1. Then if x1 < x2 then fx(x1) is < or = fx(x2) that is f is always non-decreasing function and limit of f h tending to 0 + that = fx for all x that is f is continuous from right at every point.

So now let us see that what is the relation of this capital f function with the probability mass function in the case of discrete random variable and with probability density function in the case of continuous random variable? Further we also have a point that if I have a capital f function which actually satisfies these 4 properties. If that is so then certainly that function will be cdf of random variable that is a very striking property here.

(Refer Slide Time: 42:27)

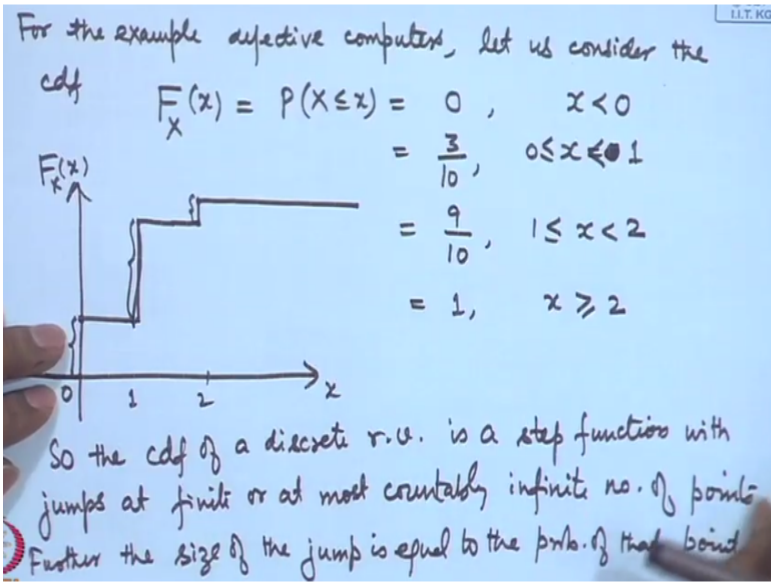


Conversely if f is a function defined on R so R to R basically satisfying the above 4 properties then it is cdf of a random variable x that means we can actually find out random variable for which this will be the cumulative distribution function. So now let us look at the relationship that it may have with the discrete and continuous random variables. So when x is discrete then see we can find out the pmf from the cdf. given the pmf we can find the cdf so both the things can be found out.

Let us see. If I consider cdf then it is probability of x < or = x that is the probability of all those points which assign the value X to < or = x. Since it is a discrete random variable say with probability mass function say px then this is nothing but probability of x = xi, xi < or = x. So that is becoming simply probability of px(xi) xi < or = x. conversely if I want to find out px(xi) then what it is?

It is probability of x = xi. Now if I have assigned my values in the space of values of capital X in the sequential order x1, x2 and so on that means we are saying like this. Then this is nothing but probability of x < or + xi - probability x < or = xi - 1. Then this is simply becoming F of xi - F of xi - 1. So given cdf one can find a pmf. Given a pmf one can find the cdf. Let us look at the example that we did here in the case of discrete random variable I defined this probability mass function. For this one let us consider.

(Refer Slide Time: 45:50)



For the example of defective computers let us consider the cdf. So let us look at this. Now from the nature of the distribution it is clear that random variable does not take values before 0 actually the values which are taken are 0, 1, 2. Now this function is defined on the whole real line so naturally this probability of x < or = x that is actually 0 if x is < 0. Now first value that it is taking is at x = 0 and the probability for that is 3/10.

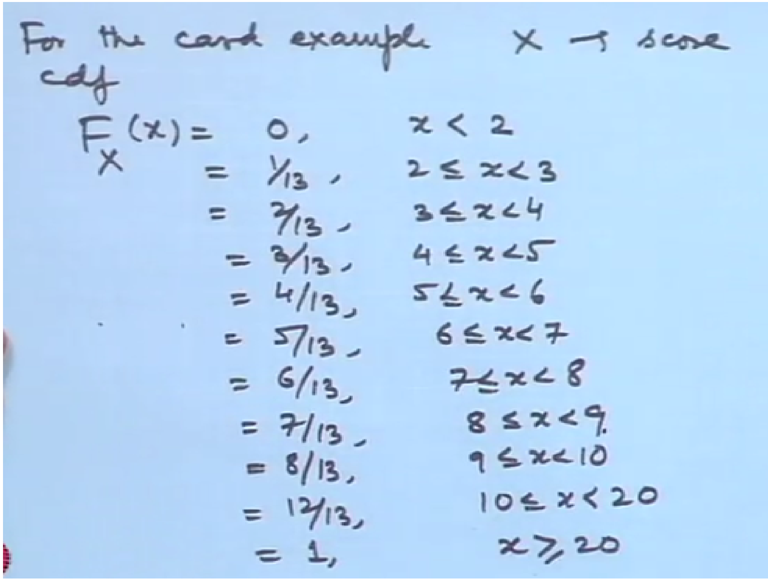
So actually you can say that it is equal to 3/10 if x = 0, but now again you observe that from 0 to 1 further addition of the probabilities are not there. So actually this statement you can qualify you may say 0 < or = x < 1 because up to 1 there is no further addition. Now at 1 you are actually having 6/10. So this probability will be added here. This will become 9/10 for 1 < of = x < 2. At 2 you are further adding 1/10 and therefore this is becoming 1.

Actually if you plot this it will look like this. Up to 0 you have 0, from 0 to 1 you have 3/10 then from 1 to 2 it is 9/10. So, something like this and thereafter it is becoming 1 so if this side we have x then fx is given like this. Now this is further revealing if you look at this function, this is actually discontinuous at 0, 1, and 2 and you look at the discontinuity and you can actually look at this is basically jump the value was 0 suddenly at 0 it is becoming 3/10.

So if you look at the size of the jump that is 3/10 that is precisely the probability of x = 0. Similarly, you look at the jump size at 1 it is 9/10 - 3/10 that is 6/10 and that is precisely the probability of x = 1 that is 6/10. Similarly, here if you see this jump size is 1 - 9/10 that is 1/10 and that is precisely the probability of x = 2.

So we can conclude that the cdf of discrete random variable is a step function with jumps at finite or at most countably infinite number of points. Further the size of the jump is equal to the probability of that point. See if this was given we can actually calculate the probability of x = 0 by taking this difference, the probability of x = 1 by taking this difference the probability of x - 2 by taking this difference.

(Refer Slide Time: 50:53)

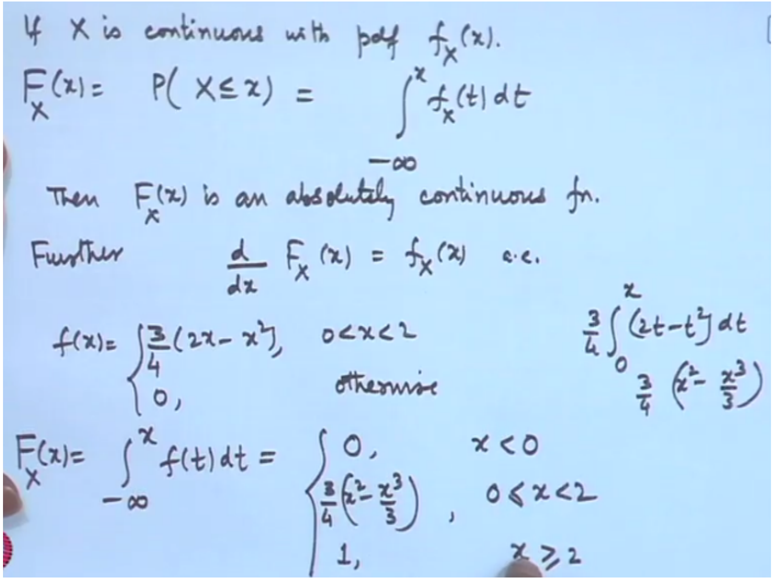


Left us further consider for the card example. So for the card example here x was a score and we look at the cdf. Now you look at the assignment here. The values are starting from 2 up to 9 then 10 and then 20. So there will be jumps at these points. So you can describe like this it is 0 for x < 2. It is becoming 1/13 for 2 < or = x < 3. I tis 2/13 for 3 < or = x < 4 and so on 3/13, 7/13 for 8 < or = x < 9.

Now at 9 you are again having the probability 1/13 so that is added it is becoming 8/13 for 9 < or = x < 10. But at 10 you have the probability 4/13. So this is actually becoming 12/13 for 10 < or = x < 20 because the next value is assigned 20 that is 1/13. So it is becoming 1 for x > or = 20. Now this is slightly lengthy example, but it is also quite revealing here. For example, you can see the jumps are at the points 2, 3, 4, 5, 6, 7, 8, 9, 10, and 20.

So for example if I want what is the probability of x = 7. So probability of x = 7 is found out by looking at the size of the jump at 7 that is 6/13 - 5/13 that is 1/13. Suppose I want what is the probability of f = 10 then you look at this at x = 10 the jump size is 12/13 - 8/13 so it is becoming 4/13. So that is actually the probability of x = 10 that is 4/13. Now if the random variable is continuous then you can se that there will be a slight difference in the description here.

(Refer Slide Time: 53:31)



If we are having if x is continuous with pdf say fx then by the definition of pdf if I look at fx that is probability of x < or = x that is integral from - infinity to x, fx or ft dt because we have seen the definition in the case of continuous random variable that the probability of an interval is given by the value of the integral of the density over that interval. So now in this case the interval is - infinity to x so integrate from - infinity to x this one.

But if we look at this carefully this is actually denoting indefinite integral of the function f and we know that if capital F is defined as an indefinite integral then this will be an absolutely continuous function F is an absolutely continuous function and further we will have d/dx of fx = fx all mostly. So let us look at the example that we have done earlier. So let us look at this continuous distribution. So here fx = 3/10 * 2x - x square for 0 < x < 2 it is 0 otherwise.

Now if you want to calculate this fx then this is by definition integral of - infinity to x ft dt. Now naturally you can see that there will be a nonnegative value only in the interval 0 to 2 in other place it will be simply because you will be integrating 0 so for example if x < 0 then this is simply 0. Now if x is between 0 to 2 then you are actually integrating 3/4 2t - t square dt from 0 to x only.

So this integral as we can easily see it is equal to x square - x cube/3. So this value is actually becoming 3/4 x square - x cube/3 and then we cross 2 then this will become the full integral and therefore this will be simply equal to 1. So this value will become equal to 1 for x > = 2. Now as you can see you can actually include 0 here or 0 here it will not make any difference because this function is actually continuous function.

See value at x = 0 is 0 from this side also it is 0 and at x = 2 if you look at this value is 1 the value at x = 2 from left hand limit and the right hand limit both are 1 so this function is actually continuous and differentiable function and therefore if you look at the derivative fo this you will get this here so this is the case in the case continuous random variable the relationship between the cdf and the pdf now in the next lecture.

I will be explaining how this concept of cumulative distribution function is quite useful when we are discussing mixed random variable because in that case we do not have to separately talk about the density function and the mass function and the cdf itself will given the complete information. We will further look at characteristics of the distribution such as expectations, moments, moments generating function, mean, variance, median, and say quartiles, measures of skewness, kurtosis etc so that we will be covering in the following lecture.