Transcriber's Name Maria Priya

Applied Econometrics

Prof. Tutan Ahmed

Vinod Gupta School of Management

Indian Institute of Technology, Kharagpur

Module No. # 02

Lecture No. # 16

Probability Mass Function (PMF) Probability Density Function (PDF)

Hello and welcome back to the lecture on Applied Econometrics. We have been talking about the probability distribution for discrete random variable and continuous random variable. Now, we need to talk about couple of more functions. (refer time: 00:37) And these are called probability mass function or PMF, as we famously call it; mass function, PMF; and probability density function or PDF, probability density function. And this PMF is a term that we use for discrete random variable, discrete random variable.

And for probability density function, we use continuous random variable. Okay. So, we are just going to explain what is a probability mass function and probability density function. So, essentially, we we spoke about this before actually, we de without specifying the names. Essentially, when we convert the frequencies into a probability, we use certain formula to convert that, right. So, we have seen that; remember the example we have given for female labour force participation. So, we sort of created a formula, right.

And that formula actually explains, given a certain family income, what would be the female labour force participation. Okay. Now, that formula for when we use in case of a discrete random variable, we call it a probability mass function, and when we use in case of uh continuous random variable, we call it probability density function. So, we will explain this term, mass and density, how they are different, when we just do couple of examples, and we will see, uh you know, how we are actually using these functions. Okay.

So, let us do an example for uh probability mass function or for a discrete random variable, how we constitute the probability mass function. So, let us say we are talking about a coin toss; our old example, very common example, coin toss. So, again, we are tossing; so, let us say example, example 1. So, you are again cos tossing 2 coins. And when you are tossing 2 coins, you are getting the events. So, let us say probability of head equal to 1; we already know. So, that is one-fourth.

And let us say probability; sorry, probability head equal to 0 is one-fourth; probability head is equal to 1 is equal to half; and probability head is equal to 2 is equal to one-fourth. We already know that; we have, we have done this example several times. And there could be no other value; so, we can write is equal to 0 otherwise. So, you sort of talk about all the values of the variable along with their probabilities. So, this is this the idea. So, value of the variable, along with these probabilities; this value of the variable, along with these probabilities; value of the variable, along with these probabilities.

So, you basically used a formula to convert that, right. So, you basically considered that 2 coins are unbiased, they are independent events, and that is where you sort of get the probability for these uh events, right. Now, let us do another example. This is a simple one. Let us do a little complex. (refer time: 03:33) And let us say, I really do not know the probability of head occurring, okay. And my random variable here; in the second example; my random variable here is that, let us say; so, basically, you are repeating the event un un until you get the, until you get a head, okay.

So, outcome; observations until you get a head, you get a head. So, your random variable is basically; you sort of do this experiment, you keep on ra doing the trial unless and until you get a head, okay. And let us say, for this coin, since you do not know probability of getting a head, let us say is equal to p, okay. So, that is basically, uh you know, because, this time, since you do not know about the characteristics of the coin, you do not know if it is a 0.5 or whatever, okay. So, you just simply denote the probability as p.

Now, let us say uh you get a head and and then you, we will have the probability p, simple. But if you do not have a head in the first trial; so, let us say you have a tail first and then a head, so, then, in the first, in this case, your tail is going to be; so, probability head and probability tail, these are the only 2 possible outcomes; so, tail is going to be 1 minus p. And then you have a head; so, that is going to be 1 minus p into p. And you can, you can sort of write down, write it down like, let us say, this, this; you are basically indicating it as a trial, probability of trial.

So, this is the first trial, and this is the second trial, let us say; this is the third trial, okay, this is the third trial and so forth. We are just going to write it down. So, let us say, in the third trial, you are going to, you see that there are 2 tails and 1 head, and you basically have 1 minus p 1 minus p into p, okay, or or basically you can write 1 minus p square into p. In the next trial, you will have probability T is 4, and the value is going to be just following the same logic, 1 minus p cube into p, okay.

And, if I sort of, you know, get a generic formula for this, I will write; so, basically, if I have let us, let us say n number of events, T n; so, then, I will have; the value will be 1 minus p n minus 1 into p, right. And here, here, what I have done is, I have constitute uh constituted a function. You see that I have actually got a formula for us to sort of, you know, put any number into that and get an output, right. So, this is the function I have just created, this is the function I have just created. All right. Now, this is my probability mass function.

And then, I will write down, this function is valid if; so, my n can only take integral values. So, I can have sort of, you know, 1, 2, 3, 4, whatever number of trials, but I cannot have 1.5 number of trials. So, if n is equal to 1, 2, dot dot dot, whatever, and, and let us say it is up to m I have defined, and 0 otherwise, 0. I will write down, 0 otherwise, okay. So, it means that, within this range, I can, say I can allow the co the coin tossing up to n number of, m number of times.

So, then, you know, or it could be in infinity also; I can, I can allow the coin tossing up to infinite number of times. But I cannot allow 1.5 trials, because that is not simply possible. So, that is the reason I will write 0 otherwise. So, this function that that the function that we have derived, that actually represents the probability mass function. Now, going back to the previous example, if we sort of plot it; we can also plot it. So, if we plot it, we can actually see that, let us say, the probability values one-fourth, half; and I will just explain why did we plot it.

So, for 0, 1 and 2, we will have these values, this and this and this. So, this way, we can actually sort of, you know, understand the probability, the concept of probability mass function. Okay. So, these are the 2 examples we have done. Now, let us actually try to do a probability density function. Let us try to understand (refer time: 08:19) how do we understand a probability des density function. Now, we, we have said that it is for a continuous random variable.

Now, if I have a continuous random variable in the first place; so, I will see, let us say I have a continuous random variable, and I have my y here, I have x here. And usually, we know that it will look something like this, let us say a curve. Now, always, uh if I have to get a probability, I will try to get the area under the curve, let us say. So, if I have to get a value, the probability, let us say, I will get the area under this curve. And I I want to get a probability value between a and b, and I will get the area under the curve, where my 2 limiting values are a and b.

Now, I will explain to how do you obtain this actually in in when you do a problem. But before that, as I as I said, it is very important that we familiarise ourselves with the concept of probability density function, okay. So, we can also write down; I have seen this probability; write down; probability that X lies between a and b is going to be some function, okay. f x, f X dx a to b. I already said that where we explained the concept of uh (refer time: 09:41) probability dis uh probability distribution for a continuous random variable.

But here, we have to first understand (refer time: 09:48) the idea of probability density. What is a probability density and how is it different from the, when you say the it is different from the probability, how is it different from a concept of mass, let us say. Okay. Now, when we did this example, when we did this example, here, this one, let us say, when we did this uh 2 coin tossing example, we saw certain values, certain probability values for each of these uh different outcomes, okay. We, we may call it a packet of probability, okay.

Also, we had this example here, which we did previously, (refer time: 10:33) this one. And if I actually draw, if we plot it; so, let us say, I have value, this is the this is the outcomes of, this is the values of x; 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. And I have, if I get the probability values here, so, this is: 0, I do not have value; 1, I do not have value; 2, I have a value, which is 1 by 36, let us say. This is; let me use a different colour. This is 1 by 36; 3 is 2 by 36; 4 is 3 by 36, 3 by 36; 5 is 4 by 36; 6 is 5 by 36; 7 is 6 by 36.

Again, this is this one is going to be, 8 is going to be 5 by 36; 9 is going to be 4 by 36; 10 is going to be 3 by 36; 11 is going to be 2 by 36; and 12 is going to be 1 by 36. Okay. So, if I just, now use a different colour; it is look, it is looking pretty messy. So, I am going to have, for 2, I am going to have this point; for 3, I am going to have this point; 4, this point; 5, this point; 6, this point; 7, this point; again, coming back to 8, this point; then I am going to have 9, this point; 10, this point; 11, this point; and 12, this point. Okay.

Now, if I sort of draw these lines here, these scripts, okay. So, these are the probabilities for corresponding values, all right. And these, I can call a probability packet. Okay. Now, this is possible, it is possible to get this kind of distinct probability values for each of these values of the observation, d because of the reason that you have these values discretely available, right. You cannot get these kind of values for a continuous random variable, and the reasons we have explained why. Now, how you can get a value for a continuous random variable?

So, I will give you an example to actually explain this, how we can actually get the values. And here, what do we do is, we actually get a distribution, okay, we actually get the area of co, under a curve. And let us say, the example that we are giving, let us say, the temperature in Kharagpur, let us say it can vary between say 10 degree Celsius to 30 degree Celsius, let us say. And what you can do is; and let us say it is like the temperature is equally distributed across the year, so, all of this, uh you know, uh so, it is the the the line is parallel to x axis.

And if I want to sort of uh calculate the, you know, the; so, the area under the curve is going to be always 1, because it is representing the total probability. All right. So, this is 1. Now, the thing is, here, I cannot really have, I cannot really have value for say, you know, a particular set in; so, if say, let us say 18.5 degree Celsius. I cannot get a value. But what I can get, I can always get a value between 18.5 and let us say 20.5 degree Celsius, let us say. We can always get the value for an interval.

For an interval we can get a value, okay, but we cannot get a value for a specific va, specific uh temperature. Now, to sort of conceptualise it, conceptualise uh the idea of density; so, here, how we calculate it is like, the area under the curve, we know is 1; so, 1 is going to be, for the rectangle, it is going to be the base which is like, let us say 20 here; 10 to 30, 20. And I have certain height, okay, so, height is here. Now, the height h is going to be then 1 by 20, which is basically 0 point uh point, 0.05. Okay.

So, this height is known as the probability density. Now, note that probability density, the height is actually not the same as probability, okay. Because, depending on the length of your base, depending on the length of your base, your height, the value of height will actually change. And you may have the probability density more than 1 often, right, depending on what value your base has. Or it could be very small either, I mean, either both are, both are possibilities.

Now, uhm; so, you need to be mindful about the difference between a probability density and uh the probability. So, when we plot a probability mass function, we only talk about the probability, so, mass of the probabilities that you sort of get. But here, you do not get a, you get, you you do not use probability, but you use a density. It is a relative sort of idea, where you are dividing basically the this total probability by the base of of your, you know, of your distribution.

And that is the reason we have to keep in mind that, for continuous random variable, you always use probability density which is not probability, okay. And we see, when you plot a normal distribution and so forth, there we will be using probability density, okay. So, that is the idea of a probability uh density. Now, the example that I have given uh is that, there I have like continuous temperature distribution. Now, it could be like uh, uh or or rather it is a sort of, uh you know, the, there is no gradient, uh it is always the, you know, that, the that is sort of the distribution of the temperature is, across the year is same; it may actually vary also, and I will sort of give an another example.

(refer time: 16:37) So, the same example, let us say, but here, I will vary the, you know, the distribution, the weight of the distribution across the different months. And let us say, I have the, you know; Kharagpur region is mostly summer, so, let us say I will have something like this. Here, let us say it is a, let us say 10 degree Celsius to 30 degree Celsius. Let us say we have, the highest temperature goes to up to 30 degree. Now, here, what we do is, you know; uh now, given the fact that; I am assuming that the temperature is, uh is mostly summer, so, I am basically (putti) given more weight here and less weight here.

uh So, this is basically, you know, the the area basically you get. And here also, you can, in the same way, you can calculate the area of a strip. Let us say I want to know the the area between these 2. Let us say I want to know what is the temperature, what is the, you know, uh probability of the temperature be being 20 to 22 degree Celsius, let us say. Okay. So, these are all doable; it is basically a simple geometry. The example I am co sort of giving you is really a simple example. And we can sort of, you know, like, use this uh area of right angle triangle here.

So, you can, you can basically calculate, let us say, the area of the, you know; the total area of the curve is going to be half or base is basically 20; 30 minus minus 10 into some height. And if you want to sort of calculate this area, let us say I am I am first interested in this area. Let us say how many; let us say; I use; years of calculation; I use this one. So, I want to know what is the percentage of the year that the temperature is between 10 to 20. And we know how we calculate the area for right angle triangle. So, it is, here it is 10 into some h prime.

This is h prime; this is h. Now, we know that h prime and h is actually, you know, it is a some uh value, uh you multiply it by the base, right. So, it is going to be half 10 into some 10K. The ratio is going to be constant for all the different right angle triangles; it is basic class 10 math; 20 into 20K. Okay. So, you cancel K, this. So, 100 by 400. So, which is basically 25 percent. So, basically, it means that, in Kharagpur region, uh for to get, sort; in 25 5 percent uh, you know, uh percent of time uh over the year, you have temperature which is which is basically uh be between 10 and 20 degree Celsius; whereas, for 20 to 30 degree Celsius, you have the rest 75 percent, okay.

So, this is how you calculate probability for a continuous random variable. uh Now, these are the simple examples; I have used simple geometries. Of course, uh when we actually do uhm, you know, uh problems, the curves are not really that simple. So, you can really have complex curves, you know, of of different contours. So, we will see, uh either we have to use integral calculus or we have to use standard tables. And luckily, fortunately for us, we have many standard tables like z, for z-distribution, t-distribution, which we will be using.

uh We do not have to really do uh (calc) uh ma use a calculus; putting it, simply draw these uh values from these tables, and we will see how we can do that. Okay, so, with this, we kind of, at the end of this lecture, and last, (refer time: 20:17) I will just talk about one more concept, and that is cumulative distribution function or CDF. So, we have learned PMF, we have learned PDF, and now we are talking about CDF. Now, as the name suggests, the cumulative, it is quite intuitive, okay.

We should be able to make sense what we mean by cumulative distribution function. Now, basically, you add up all the probabilities. You know, for a given value of the random variable, you see, wha you know, like uh, what is the sum total of the probability for that random variable before that particular value. Let us say I want to know the, you know, uh CDF. So, I will define CDF as like this. So, let us say CDF is going to be something like probability of X less than a, okay.

So, all the values are adding up to a. And we sometimes represent it as or we write, it is a function; this is also represented as function of a. Okay. So, it it, the moment I write this, uh the moment I say that this representing the CDF, so, essentially, it means you have you have kind of summed up all the values up to the value a. Okay. And this is actually equally true for both the, uh you know, uh PMF as well as PDF, or continuous random variable as well as the discrete random variable.

So, you know, uh if if I write, let us say, probability of X, X less than equal to 25, you will simply write, is equal to, basically, F of 25. This is the same thing. Now, I will ask you a question. Okay. So, I have been talking; na now, it is my turn to ask you a question. And let us say I ask you, uh for a continuous random variable, what would be this value, F of infinity? So, basically we know, when I write F of something, that is basically the CDF. Okay. So, if I ask you what is the value of F of infinity, what would be your answer?

Pause for a moment and think; and then, I will answer it. I hope you got the answer. So, this is going to be equal to 1, because, the moment I say infinity, I am going from, you know, minus infinity to all the way up to plus infinity. So, I am basically covering up all the possible values and the; basically, I am covering up all the, you know, probabilities for all the possible values between minus infinity to plus infinity. So, that means, it is always going to be 1.

And if that is the logic, then what is going to be for this; what is going to be this value, F of minus infinity? So, if F of minus infinity, if you want to think for a moment, F of minus infinity is basically, you have not yet started; basically you are in minus infinity, you have not added up any number, right? So, that means, it is going to be 0. Okay. So, this is something you need to remember, that F of infinity uh is equal to 1 and F of minus infinity is equal to 0; of course we are talking about a continuous random variable.

Now, given this, how we can write the, uh you know, a continuous random variable, you you know, the distribution of continuous random variable using the idea of CDF? So, we can write, it is always going to be 0, F of a to infinity, okay. This is something we can always write. Okay. And as I said, this is, it is same for uh discrete random variable as well. Now, if I say that; let us say; again, let us go back to the example of this coin tossing, (refer time: 23:59) this one.

And here, I ask you that, you know, I ask you F of 1; so, which means, I am asking you, head could be 1, maximum; so, that is the maximum. So, it means, when I write F of 1, let us go back to where we are writing. So, here, F of 1, (refer time: 24:21) F of 1 is going to be what? Probability of head is equal to 1 plus probability head is equal to 0, right. So, if I sort of write down the probability values, I know probability head is equal to 1 is half and this is one-fourth. So, together, they will make 3 by 4.

So, that is going to be my; uh using the CDF, I can get the value. Okay. So, this is basically the idea of CDF. And uh we may use the concept CDF time and again. So, this is basically the basics of uh probability distribution for discrete random variable and continuous random variable. And in the next lecture; so, this is, we will we will end here. And in the next lecture, we are going to talk about expectation rules, variance and covariance rules which we are going to use uh going forward; and what is an expectation, most importantly we sort of need to familiarise ourselves with.

With that done, we will be now, we will; after that, we will be talking about different types of distribution, standard distributions that we will be using all throughout all lectures. So, with that, we end the lecture here. Thank you.