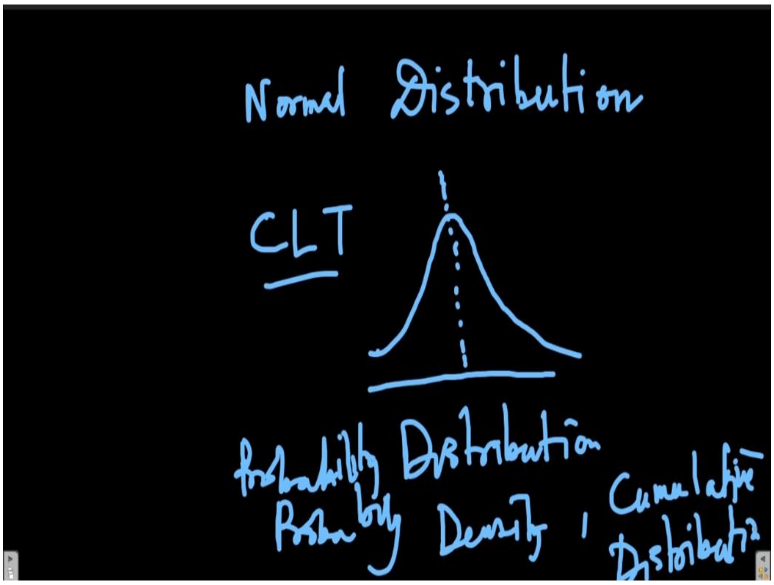
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 3
Lecture - 30
Normal Distribution

Hello and welcome back to the lecture on Applied Econometrics. We are into the third week of our lecture.

(Refer Slide Time: 00:31)



In today's lecture, we are going to cover one of the most important topic, and this is about a distribution, and this distribution is known as normal distribution. Now, why it is normal? Because perhaps it is the most naturally occurring distribution that we will see; it is all pervading; everywhere we see this kind of distribution. Now, why one particular type of distribution would be all pervading? Why it has so much of ubiquitous presence?

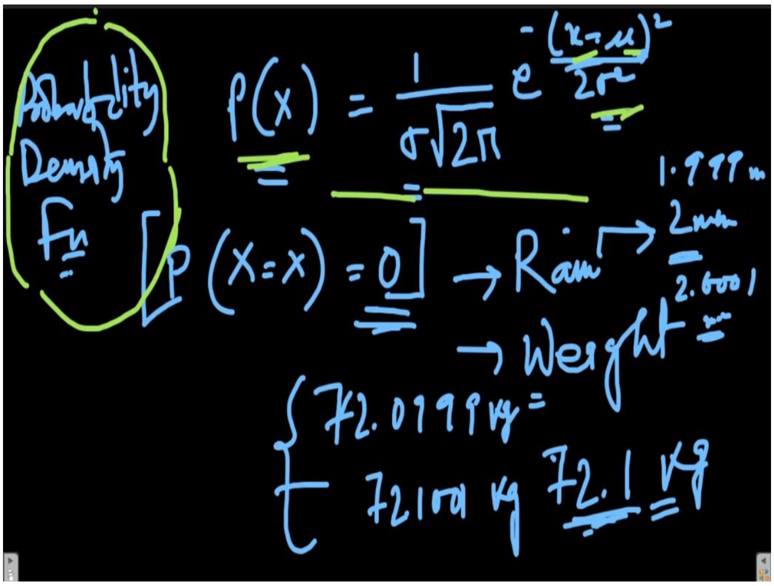
To explain that, you may have to go back to something that you have learnt already, and that is a central limit theorem, if you remember. So, in central limit theorem, we said that, when we draw a large enough sample size and when we draw them independently, the random variables are independent, and if they are identically distributed. And if we sum them up or if we take a mean of that, the mean of the distribution of the sum of the distribution will actually follow a normal distribution.

Now, that is how nature works. And that is precisely because of that reason, we see the normal distribution to be so all pervading. So, it looks like this kind of bell shaped curve, we have explained previously. So, this distribution, you will see almost everywhere; not in all the cases, but it has kind of universal presence. Now, in this lecture, we are going to talk in detail about the properties of distribution, and how it comes and so forth, how we deal with this kind of distribution and all.

But before we get into that, it is important to learn some of the related concepts which we need to explain this distribution. And these concepts, we have already learnt. So, for example, probability distribution or probability density for that matter, which is very important; cumulative distribution. We have already covered these terms; I am just writing it down because, going forward, we have to use these terms.

Then we have to learn about; we already know about the area under the curve, the way we get the probability for a continuous random variable by using area under the curve, and how it is different from the discrete random variable. That, we have discussed previously, when we talked about PMF and PDF, probability mass function and probability density function, or different types of probability distribution. In this lecture, we are going to see the application of that.

(Refer Slide Time: 03:06)



Now, when we talk about normal distribution, we usually define the probability density function to be something like this. It looks a little daunting, but I will explain that. 1 by sigma; 1 sigma is the standard deviation; root over 2 pi; and then we have e power minus of x minus mu square by 2 sigma square. That is basically the expression of the probability density function.

Now, I will talk about this mathematical expression in detail. And but before that, we need to talk about that what is the probability density function? what is the probability distribution function? and what is cumulative distribution function of course? So, we should not confuse these terms and we should have some clarity about that. Just note one thing here. In this formula, we are using the standard deviation; this sigma; and at the same time, we are using the mu, which is the population parameter or this is the mean, and sigma is the population standard deviation.

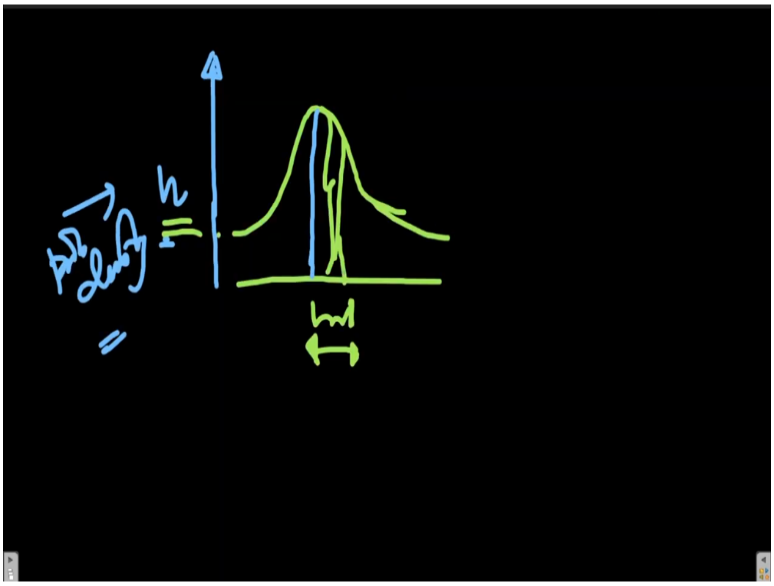
So, we are using these terms. So, we will see the relationship of this mu and sigma with a probability density. Now, we know that for any given value X, for a continuous random variable, the probability of X = X is equal to 0. So, for a continuous random variable, for any given value of X, you cannot have a probability value. And the reason is that; well, so, if you take examples of, let us say rain, or if you take example, let us say height or weight; anything you want, you can take.

Now, I say that someone's weight is, let us say, maybe 72.1 kg. Now you measure his weight. And you cannot precisely say that his weight is 72.1 kg, and I have explained this before. He can be like 72.1001 kg or 72.0999 kg or something like that. So, it is almost impossible to get the precise 72.1 kg, taking into account every other, like, as if he is the constant weight; I mean, perhaps, depending on his breathing exercise, his weight might be a bit more or less.

So, similarly for rain, you cannot really say that it rained only, let us say, 2 millimetre. So, it is very difficult to reach to the exact number. It could be like 1.9999 millimetre; it could be like 2.0001 millimetre or anything. Now, because of this fact that you cannot really reach to a very specific number when it is a continuous random variable, we cannot have any probability value.

So, that is why, for a given value of X, the probability is always going to be 0. Now, and that is the reason we talk about; let me use a different colour; so, that is why we talk about this probability density function, which is not probability. And the probability density function, if we draw a normal distribution or any other distribution;

(Refer Slide Time: 06:57)



So, let us say the normal distribution, we use probability density here. And here, the probability density is going to be the height of the distribution. So, this side, the y-axis is actually representing the probability density. So, this is the probability density; this is not probability. Now, probability density; now, for continuous random variable, a probability distribution function is basically the probability density function.

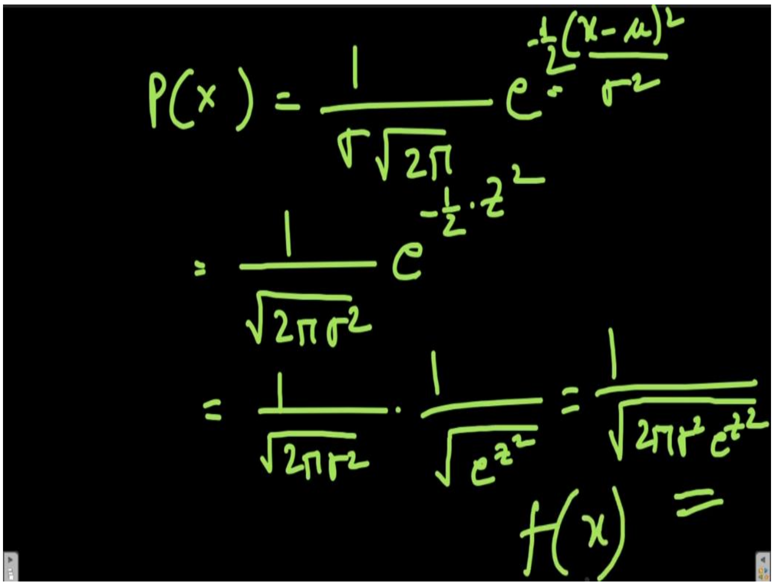
And so, this height, as I said, is just the height; it does not talk about the probability. So, if I impute value of x, mu and sigma here, we will get some value of P of X, and that is the probability density. So, that is basically the height. If we impute the value of x, sigma and mu, we will get this height. So, this is basically, what we are getting is the probability density; not the probability. Now, how do we get that?

So, we know that we have to get the area under the curve; we have already seen when we talked about continuous random variable, how to get the probability. So, basically, we use this height, and then we use the value of the variable or the span or the window for which you are basically; the range for which you are taking the value of the variable; and you integrate that. So, we are going talk about that again.

Before we go there, as I promised, there I am going to explain this equation in detail. So, this is a little daunting equation; I will be trying to actually simplify this daunting equation to a very understandable form. And we are also going to use, going forward, we will be also using some Excel to be familiar with what goes on here and how to easily handle the whole thing.

So, there are easier way to actually handle the whole thing. We do not have to integrate this kind of function; there are simpler ways. So, let me just explain this functional form a little bit.

(Refer Slide Time: 09:16)



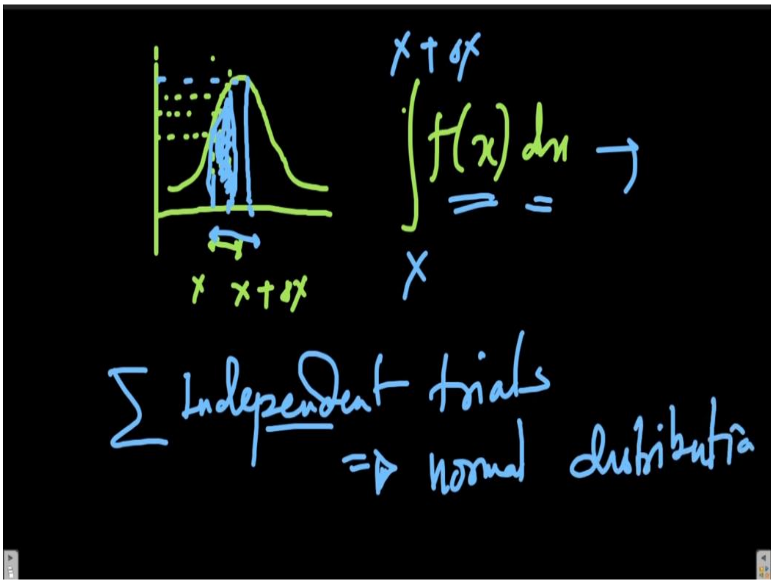
So, as I have said, let me write it down again; probability of X is equal to; it is a good idea to actually remember this, but there is no harm if you do not. So, root over 2 pi; and I am actually going to tell you a simple way to remember this; e to the power; x minus mu square by sigma square. So, same; we initially have written 2 sigma square; and here I am writing, basically I am taking half out here.

Now, what is the simplified form? How we can simplify this? Let us see. So, this sigma is the standard deviation. Now, if I bring the sigma under the square root, I get that variance. So, I have no problem in doing that. So, 2 pi sigma square; somewhat easier. Now, we have everything under the square root. Now, e to power minus half; so, that is important; that is interesting.

Now, I will represent this x minus mu by sigma into something called Z; and I will explain what this Z is later on. So, let me write it down; half Z square. Now, when I have, use a simple way we deal with the powers, when we have e to the power minus half, how do we write that? It is basically 1 by root over of e, right? So, this is basically 1 by root over of e. So, if I do that, so, it will be root over 2 by sigma square; and this is going to be 1 by root over; so, this is, because of minus sign, it is going to be 1 by; and because of half, it is going to be root over e to the power Z square, e to the power Z square.

And now, if I take all the things together, so, it is going to be simple, one term; 1 square root of 2 pi sigma square e to the power Z square. So, that is basically the functional form. So, we can remember just this one; this is enough for us to deal with the normal distribution. So, this is the probability density function and a simplified way of expressing it is this. Now, I said that we are going to actually talk about the area under the curve, and we are going to actually explain in terms of the probability density. So, let me do that.

(Refer Slide Time: 11:47)



Now, when I have, let us say, for this curve, I said that; so, if I have the axis here. Let us say this is the point, this is the height I get; and for all the different points on the normal distribution curve, I get different heights. Now, when I actually; now, if I represent; so, my previous equation here; so, if I do for our convenience, if I write down as F x; so, if I basically try to get the area under the curve; so, basically the probability.

For probability, what we do? We actually integrate. So, F x the probability density over dx; over this area, this interval, we integrate that. Now, if it is, let us say X and this is X plus del X. So, how we write it? We write it basically, here we write X plus del X; from X. So, this is basically, this indicates the probability under this normal distribution between these 2 points. And how we do that is basically, if it is kind of the height and if it is kind of the base; so, you can say that; you take this area here, of this rectangle or you can say trapezoid; so, it is basic calculus.

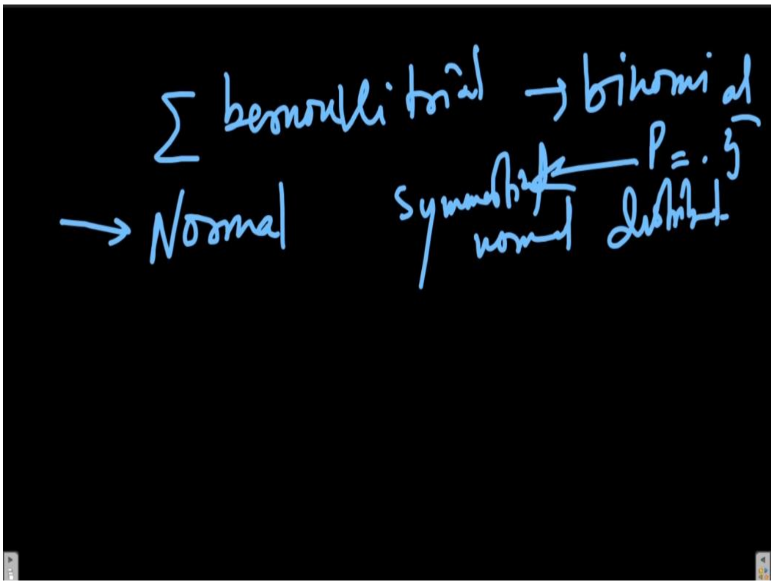
So, you basically multiply this base and height; and then you sum up all these little strips; and then you get the total area under the curve. So, this is how we basically calculate the area under the curve. Now, we can also interpret it in a different way. So, if we consider only the probability density, so, which is basically the height; and I can never take the area of this height; so, it is only a line.

Now, for this line, the width is basically 0; there is no width essentially. And, so, that is why we can never take the probability for any specific value, because, when we use a specific value, we can only get a line, only get the height; and the height is without any width. And when the width is basically 0, so, you cannot get any area. So, that is why, for probability density can never be is equal to a probability. This is something you need to remember.

Now, so, essentially, again going back to the central limit theorem that we spoke previously, the basic theorem, central limit theorem and why that actually leads to a normal distribution. Now, if we think like this, here again just recollecting the central limit theorem, if we think that, there are lots of independent trials; so, basically, drawing sample is essentially, doing some independent trials, and you do not know the original distribution; so, you do all these independent trials.

And when you do these independent trials, you basically have the probabilities, all these small probabilities, and you see here that you always integrate to get this probability. So, essentially, you sum up these independent trials. Now, when you sum up the independent trials; to get the probability, you need to sum up. And when you do the sum up, essentially, it leads to some normal distribution, because that is what we have seen in the case of, when we used Bernoulli to binomial.

(Refer Slide Time: 16:08)



When we sum up all these Bernoulli trials, we reach to something called binomial. And we have seen the binomial could be approximated to normal; we have seen that previously. And if we have lots and lots of trials of binomial, it will be more and more normal. So, as we increase the number of trials, it is going to be more and more normal. And if we have the probability of success in binomial is equal to 0.5, it is going to be more of a symmetrical normal distribution.

So, we have seen that; it leads to symmetrical normal distribution. So, that is basically the reason. This is one example like Bernoulli to binomial normal; but any other distribution you can take. When you sum them up; we have seen that; when you sum the IID random variables, that essentially leads to our normal distribution. We have explained in detail in central limit theorem.

Now, we are going to talk about little bit in terms of the empirical part of it, and that we are going to cover in the next lecture. So, let us divide this lecture into 2 parts. So, in the next lecture, I will cover the empirical part of the normal distribution and we will illustrate. With this, we end the lecture here. Thank you.