Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

Indian Institute of Technology, Madras

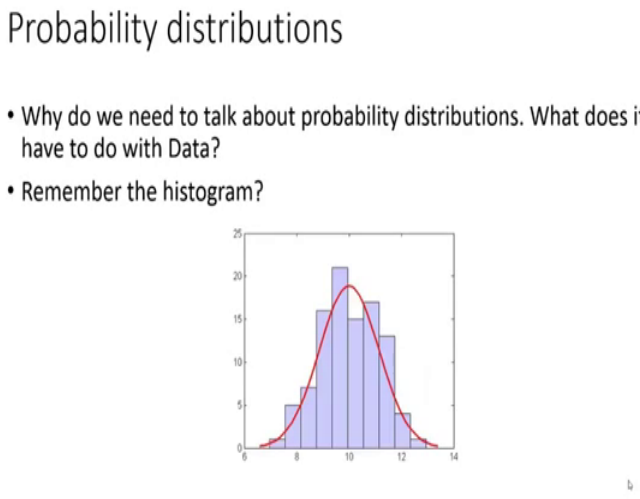
Module - 02

Lecture - 06

Random Variables and Probability distributions-1

Hello and welcome to the next lecture in our course Introduction to Data Analytics. In this lecture we are going to be talking about Random Variables and Probability Distributions and this would be the first lecture of the series that cover this topic. Just a recap on what we have completed so far, we finished looking at Descriptive Statistics and the use of various use of graphical and visualization techniques in descriptive statistics as well as the use of summary statistics. Within summary statistics, we looked at measures of centrality and measures of dispersion.

(Refer Slide Time: 00:54)



So, jumping into this topic, quick question is why do we need to talk about probability distributions. What does it have to do with data? It is just like a mathematical concept, why, what does it have to do with data. And the quick answer to that question is, if you go back to the use of the histogram we express that as a way of describing data.

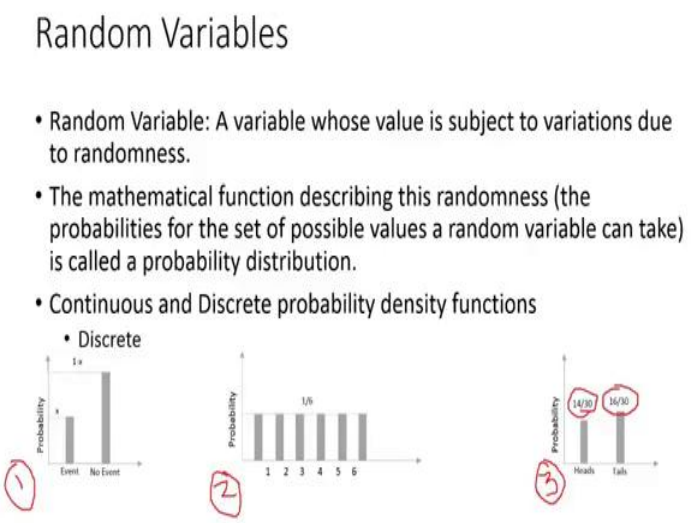
Essentially the histogram is, if you look at this picture on the slide, the histogram is those vertical gray, grayish blue bars that you see on this graph and that describes the data that summarizes the data in some way.

But, you might be of the belief that if you re do this exercise, if you collect a new sample you will get bars that looks slightly different and the question is, is it really coming from a probability distribution, is it coming from some other mathematical function that closely approximates this histogram that you are seeing. And that red line that you see on this is the attempt to fit this mathematical function and the core idea here is that, this data is being generated by this probability distribution function, which is that red line and the histogram is, what you see in terms of data.

Because, not every time you going to get data that looks exactly like the red line, so it is in this context that you can think of a probability distribution also as a way of just describing your data. But, I would say describing and not summarizing, because it is fairly comprehensive, it just does not give you one number or one thing. It gives the full form and shape of that data and you can think of it as an exercise also in modeling your data.

So, you are not just describing it, you modeling it. So, in that context probability distributions are very important and we will also see how in various other things, not just describing data, but even in terms of doing more advanced analytics in the machine learning parts in the statistical inference parts, the use of probability distributions is critical. The think of a data set has random numbers that are being generated in accordance to some mathematical function is the whole idea behind, the use of probability distributions with respect to data.

(Refer Slide Time: 03:26)



To do this, we kind of have to understand some basic concepts, which is and the first basic bond is to understand what random variables are. Random variable is essentially a variable whose value is subject to variations due to randomness as a post to variations due to some other phenomena. So, we are all familiar with the concept of constant, which just means it is a fixed number and variable. Here I am talking about the variable that you probably learnt in algebra in high school, non random variables and there you learnt that the variable is essentially something that can take on many possible numbers or any possible number.

But, the distinguishing factor between a variable and a random variable is that, with a regular variable once you fix all the externalities, then the variable takes on a specific value. So, let me give you an example. So, you take something simple like, force is equal to mass times acceleration. So, force is equal to mass into acceleration and you might say, all these are variables and they are and that is true. So, force can be any possible value, given what I have just told you, force can take on various numbers as it is value with some units.

But, once you fix mass and you fix acceleration, force will take on a very specific value. As oppose to a random variable, where even if you fix all the externalities, the best way to describe the random variable would be to say that it can still take on a set of possible values and that set could be a very large set, it could be infinitely large set. But, the

variable itself can take on many possible values and each of those values have a specific probability associated with it and beyond that, you are not going to, you cannot reduce the variable beyond that by definition.

Even after you fixed everything around this variable, you still have to describe the variable with a probability states phase. So, let us, the best way to again get this even deeper, to understand this even better is to talk about somewhere you specific probability distributions and that is what we are going to do in the next part of this class. But, before I proceed I just wanted to tell you that, I am broadly breaking up the idea of probability distributions into discrete probability distributions in continuous distributions and that is just to give you some structure into it.

Those words might not make immediate sense to you right away, but what we are going to do right now is to look at discrete probability distributions. You might also notice that I am using the word probability density functions and you might not know over that word means yet, but very soon we are going to be talking about that is well. So, great, so we are going to look at the most simplest discrete probability distribution and this, it is really simple, because I think we all used it in some sense in our daily life.

So, we are here, so let us look at the first example, I will give these numbers just show that going forward, it is clear. So, we are looking at number 1. In number 1, matches the closest with a colloquial use of probability, chance, likelihood and so on and we are saying something simple, which is that the probability something happens is x. So, you might say the probability that rains today is 10 percent, is the 10 percent chance it is going to rain today.

What goes unsaid is that, there is therefore, a 90 percent chance that it does not rain today and that is what is captured in this graph. So, we more used to saying, this is 30 percent chance there it is going to rain, this is 20 percent chance that there will be an accident, there is a 10 percent chance that the product that I am manufacturing is not fit to be shift. But, essentially we are talking about these kind of binary events, where one of the possible outcomes is x and therefore, by definition the remaining possible is just 1 minus x or if you thinking of it in terms of percentage, this is the 100 minus x.

Again a very simple example of this could also be something like this, there is a 50 percent chance that, if I toss this coin I am going to get a heads and what goes without

saying is therefore, that there is a 50 percent chance that you would not get heads. In this case, that is called tails, so great. So, that is one very simple conception of probability and this is a probability distribution, it is called Bernoulli distribution, but we can move to multiple outcomes.

So, if you look at number 2, what we have there is, what you get when you role a dice. So, you role a dice and a dice has six faces and on each face you have a dot. So, you have, if you role a dice you either get a 1 or a 2 or a 3 or a 4 or a 5 or a 6 and the idea is that the probability of each of these is one sixth, if it is a fair dice and so, that is a different kind of a probability distribution. We will soon learn in our next class that is called discrete uniformed distribution, because they are all the same probability, but the possible outcomes going back to our definitions is set.

So, the possible outcomes possible values are 1, 2, 3, 4, 5, 6; the probability associated with each of those possible values is one sixth and one thing that you might have noticed by now is, if you take all the possible values and you take each probability and you add the mole up, you always get 1. So, in the case of 1, we saw that if the probability of it raining was let us say 30 percent or let us say the probability it rains today is 30 percent and the probability it does not rain therefore, becomes 70 percent. You add those two up you get a 100 percent or you get 1.

Similarly, you have six possible outcomes when you role a dice and there is a one sixth chance of each of them happening and six times 1 by 6 is 1 and the intuition for this should also be obvious, that if you role a dice or if the day passes means something has to have happened. So, you have had to have gotten one of those six numbers or you know it either rained or it did not rain, but as long as you have comprehensively covered the universe of possibilities, then something needs to have definitely happened within that universe.

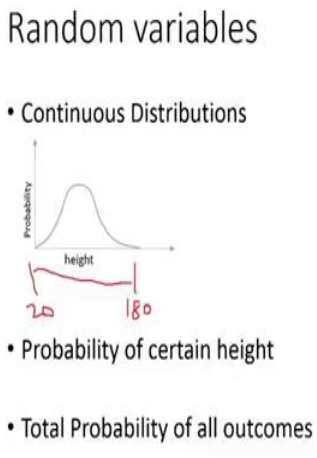
So, therefore, that should also been intuition us to why, that the probability distribution sum to 1. What we have in number 3 is the idea that, again it ties to this notion of probability not just being a theoretical exercise and you might actually have some data and you might choose to define the probability distribution based off of what you see in the data. So, if somebody came to you and said look, I do not want you to assume that, so I wanted to take this coin and I wanted to describe this random variable, which is the

probability of getting a heads or a tails and that is the random variable and I do not want you to assume, there is a 50, 50 chance.

So, you might say fine, I have nothing I cannot assume anything and you might toss the coin a few times. So, you do a data collection exercise, where you toss the coin 30 times and you notice that you get 14 heads and you get 16 tails and for whatever reason, if you do not want to assume anything about the distribution and let us say, you also do not want to do any statistical inference, again a topic that we will cover soon.

You might just be contained and saying, I am going to describe this random variable with the actual data that I see. So, I am going to actually say that a 14 out of 30, there is a 14 by 13, because you actually got 14 heads when you toss the coin 30 times. So, I am going to actually say 14 by 30 is the probability of getting a heads and 16 by 30 is the probability of getting a tails. So, there is nothing wrong with doing something like that, we would have to see if that was actually made sense do, but nevertheless if you said I just wanted to take data and I wanted to describe a probability distribution with the data that I see, then you can definitely define a discrete distribution in this way.

(Refer Slide Time: 12:57)



And now, we go on to something that is a little more complicated, which is continuous distributions. In the previous case, ((Refer Time: 13:04)) in the discrete distributions and let me erase this, all the ink. So, in the discrete distributions, what made as discrete, were that the possible outcomes would discrete. So, it was either an event or a non event, so

that is discrete, there is no half event. So, there is no half event, not there.

Same way are here, you either get a 1 or a 2 or a 3 or a 4, this set of possibilities that is the here x axis essentially has some countable number of possible states and so, you cannot get a 1.5 when you roll a dice and similarly, you cannot get a half head, half tails when you toss a coin. So, that is essentially the concept of it being a discrete distribution and with continuous distributions; however, that is not really true. The idea is that the x axis are here, so it is the same idea which is the possible outcomes are on the x axis, same thing that we saw on discrete and the probabilities are on the y axis.

Case of this is the same core concept of describing the distribution, but here the x axis is not discrete. So, what is that mean? What it means is, you take something like the probability of a certain height. A height can be a 130 centimeters, so that could be one number out here, but it could also be 130.001 centimeters, it could also be a 129.999 centimeters. So, there is no inherent discretization. You might turn around and say, look what if I had a measuring scale that could only measure in 5 centimeter intervals.

So, the way I mean or I can only measure up to a centimeter, I cannot measure less than a centimeter, because the scale that I have does not have more resolution and that is fine. You know, if your resolution for measurement is still accurate, meaning that anything between 130 and 131 gets called a 130, because of the inherent resolution of the scale. That is fine, you can create a discretized version of it, but the idea is this nothing about height or any, essentially the measurement of space.

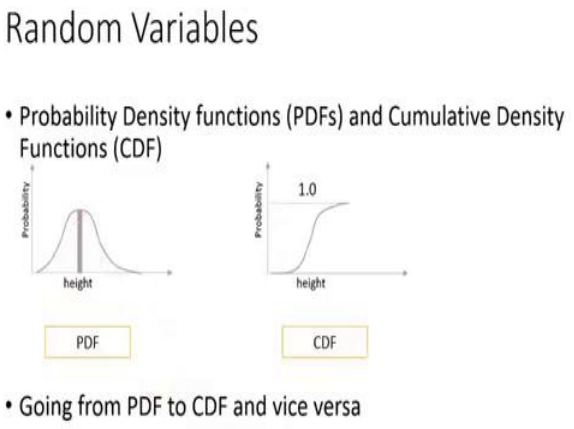
There is nothing about height that is inherently discretized, like the measurement of a dice is inherently discretized. You cannot possibly roll the dice and get a two and half, whereas if you had a five and half measure of height, you could get any possible value within the certain range. So, you might have the lower end of this being 20 centimeters and the upper end of this being a 180 centimeters. But, essentially any value between it is possible and we kind of spoke about the same concept when we spoke about discrete and continuous variables. The same concept of discrete and continuous variables applies to discrete and continuous distributions.

Now, again another important thing to note is just like in the discrete distributions, where we said the sum of all the probabilities for each possibility should add up to a 100 percent or should add up to 1. Here you cannot have a countable number of possibilities,

so you cannot take each possibility. Take the probability of that and added to 1, just because there are infinite number of them. So, you basically what you do is, you take the entire interval and sum the probability within that interval and the best way to do that is to look at this area as a whole.

So, this way when you look at this area, it is like you are taking all the possibilities within that area and summing all the probabilities for each possibility and when you do that, you will be getting 1 or a 100 percent.

(Refer Slide Time: 17:39)



So, now let us now that we understand both continuous and discrete random variables. I just briefly talk about the use of probability density functions and cumulative density functions. So, far what I have been graphically showing you, have all bend probability density functions. For each probability density function, there exist a cumulative density function and so, I will describe it in the continuous case and that is the easiest and I separately do that for the discrete.

The idea with the PDF is that, like we said for, because there are infinite number of possible states, you really does not really make sense to ask the question, what is the probability at a given point. It turns out that the answer to that question is that the probability of that given point is zero, although there is some y, there is some height for that given point. Because, there are infinite such points out here, because there are technically infinite such points. As a result, you can only say what is the probability of a

given area, so you can say I want to look at this area and you can get the answer to that question, you can get the probability of this given area by just measuring the area under this curve and that is what I have done with the gray line on this graph as well.

But, the idea behind PDFs to CDF is that, the CDF describes the cumulative probability up to a certain point. So, if I would ask the question, what is the probability within this area, you could answer it from me by looking at the area under the curve. If I ask the probability density function at this given point, you can use the function to figure out what this value is, but the probability itself is zero. But, the cumulative describes the probability from zero, from the lower end of this axis and that can be zero or that could be something like minus infinity or it could be some other value.

You know this starting point essentially, from that starting point all the way up to the point of interest. So, this is x, the PDF describes the height out here for x, the CDF describes… So, this is your point x, the PDF describes the height of this curve at x and which is y, this entire curve out here this curve that is here on the left hand side is the PDF, whereas the CDF describes the area to the left of this point x and that area is the CDF.

So, this graph is nothing but, the same as the graph to the left, where at each point x you are looking at what is the area to the left of x on the PDF and that is what you are plotting here and that should logically be equal to 1, when you complete and the reason for it is the following. We already discussed in the previous section is to how the overall area under this curve. The overall area under this curve is equal to 1, correct.

We discussed that if you take all the possible states and add the probabilities of all the possible states, ask to have you are getting up, you would get a probability of a 100 percent or 1 and. So, the CDF is nothing but, this description of the area to the left of the curve and when you reached your entire set of possible heights or in this particular case heights, but it can be anything else, then the CDF hits at the one mark when it ends.

And it is, it would be helpful for you to know that therefore, the easiest way of getting from a PDF to CDF and a CDF to PDF is, from a PDF to CDF you would essentially want to integrate. For those of you, who used integration in high school or if you heard of the term integration, that is that symbol that looks like this and the idea is that an integration covers this core concept of area under the curve. So, if you integrate up to a

point x, so let us say just for as an example that this started at zero, but you could change that.

It could start y, it could started minus infinity, but you essentially if you integrate the area under this curve to get the CDF. You would just want to integrate from 0 to x, some f of x and the f of x here is your PDF function and that will give you the CDF and the idea of going from CDF to PDF is exactly the opposite, which is to differentiate this CDF and that will give you the PDF. So, we will conclude this lecture with that note and starting from the next lecture, we will be talking about some actual probability distributions and we will go through atleast the most popular ones in that lecture.

Thank you.