Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

Lecture - 12

(Video Starts: 00:14)

In this discussion you will be introduced to the concept of probability distributions specifically discrete distributions and their probabilities. First, we will talk about random variables and probability distributions. If you recall an experiment is in any process that leads to one of several possible outcomes. So, the random variable is just a mathematical way of describing these outcomes.

And when we plot the probability of all the random variables on a graph then this graph is what we call probability distribution. So, the random variable and probability distribution are two related concepts and we learn about both of them in much more detail. The next concept that we will see is that of expected value and variance. This is very similar to what we calculated in the form of mean and variance for a population or a sample.

The only difference being that you learn how to calculate the same for a probability distribution Expected value is generally used to find an anticipated value for an investment at some point in the future and is popularly used in industries like insurance and card games. We will see some interesting applications around expected value. Apart from that you will also learn about the important discrete probability distributions that might come across in your day-to-day applications more specifically the binomial distribution.

Binomial distribution could be used to mimic experiments like public votes in an election survey or analysing new products response via a limited launch.
Random variables part 1. I am quite sure that you have heard this well-known saying about casinos that, the house always wins what it means is that the casino gambling machines and games are designed in such a way that in the long run casinos always make money.

One or two people may end up winning large sums of money but the machines are designed such that the rest of the people collectively lose more. So, how do the casinos ensure that the house always wins it is quite simple, they use probabilities .In fact casinos use a very basic concept of probability to ensure a win for themselves, do we wonder how to understand this. Let me tell you about a simple gambling game we can play.

We can take a bag and fill it up with three balls two red and one blue each participant had to take out a ball, note its colour and then put it back in. This process was followed until each participant has taken out a ball from the bag four times. A participant who got a red ball all four times would receive 150 rupees from the house attractive offer is not it? However, the participant who got any other result would have to put or pay 10 rupees.

So, would you play or would you like to play this game is this the kind of game in which we the house should always win in the long run or are we going to incur some losses in the long run. In order to understand all this we will approach the problem in three steps. First, we will see what all kinds of combinations can come up. Then we will see how likely each of these combinations that is what their probability is.

Then we will see how we can use these probabilities to estimate the profit and loss of a player he would get on playing the game once. Let us go through the various outcomes possible after a participant has tried the bet. As you would guess the one outcome our players would want the most is that all four balls end up being red. However, that is not the only possible outcome they could, for example get three red balls in a row only to find that the last ball is sadly a blue one would not that be sad.

So, can you find out all the possible outcomes. So, let us first list down all the possible outcomes one by one we could get four blue balls, there is only one outcome in which this happens. We could get three blue balls and one red ball this could happen in 4 ways RBBB, BRBB, BBRB and BBBR. We could also get 2 blue balls and 2 red balls. This could happen in six ways. Also we could get one blue ball and three red balls in four ways.

And there is only one way in which we could get four red balls. So, in total, there are 16 possible outcomes.

Random variables part two, let us go back to some of the original questions that we were concerned with starting with how likely is the house to win or what is the profit or loss you should expect while running such a business. However, currently we cannot answer these questions and in order to answer them we need to analyse the outcomes in terms of probabilities.

For example, if the probability of getting 4 red balls in a row comes out to be extremely low. We may say that the house is likely to win on the other hand if the probability of getting 4 red balls is comparable to the other possible outcomes. It may be profitable for the players, let us think about how we can do that. So, there are various ways in which we can convert these outcomes to numbers or in other words quantify them.

Let us assign the value of the number we get after quantifying to a variable X. For example, let us quantify the outcomes by using the number of red balls. So, instead of saying that the outcome is blue red red red we can say that we got 3 red balls and X = 3. Similarly for the outcome blue red blue red we can say X = 2, as you can see this number used for quantifying outcomes is really important in statistical language it is known as random variable and is denoted by uppercase X.

It is basically any variable that converts an outcome to a number. Another important question is are there more ways in which random variable X could be defined. We can define X as the number of blue balls that we drew from the bag or the number of red balls minus the number of blue balls that we have drawn from the bag. So, the question that now arises is which is the correct random variable. Out of all these options that you have which one should you choose.

Well, the answer actually depends on the information you are interested in. In our example, we are interested in the number of red balls. Recall that if the player draws four red balls he or she wins 150 rupees. However, if the player draws even 3 to 1 or 0 red balls he or she loses 10 rupees. So, let us define X this way X equal to number of red balls. Now if you want to find whether we will actually make money or lose it we need to know how likely the different values of X are.

What is the probability of each value we will see that in the subsequent discussions.

Probability distributions. Let us see the different values of the random variable X as we defined before, it will take for all the possible outcomes of our game. Remember that X was defined as the number of red balls that we have drawn from the back. So, if the possible outcome that we look at is the 1 with 4 blue balls obviously X takes the value 0 for all.

Four outcomes where we only drew one red ball out of before X will take value of 1. For the six possible outcomes where there are two red balls and two blue balls in the various combinations, X will take a value of 2. For these four outcomes where we have three red balls and one blue ball, X will take the value three and for the last outcome where we have four red balls, X takes the value of 4.

So, basically, by doing this we have brought our 16 outcomes into 5 groups where the random variable takes values from 0 to 4. Now if you want to find out whether we will make or lose money in the long run we need to find out the likelihood of each of these values occurring, but how to do that? Let us say that we conducted experiment with certain 100 individuals, who played this gamble.

Based on each person's response we put a small note on a chart for example person A got three red balls so we put a note with his name on the chart or value 3. Similarly, person B got two red balls, so we put a note with his name over the value 2. This way as more and more people play, we will fill up the chart accordingly until we have responses of all the hundred people on this chart. So, in this chart over here on the x axis we have random variable X from value 0 to 4.

On the y axis we have the frequency the number of times the random variable appeared in our experiment, so this chart is called a histogram. So, the heights of these bars in this plot correspond to the number of times the frequency of that random variable appeared in our experiment. For example, 2 out of the 100 people drew no red balls, 10 out of 100 people drew 1 red ball. And so

on. So, recall that the probability of any outcome is the number of favourable outcomes divided by the total number of outcomes.

So, let us find the probability that our random variable X takes the value 2. The number of favourable outcomes in this case will be the number of people who drew two red balls and that number is 32 the total number of possible outcomes is 100 as that is the total number of people who participated in that game. So, the probability of drawing two red balls is 32/100=0.32%.

So, we can extend this probability calculation for all the different values of X and put all of this information in a single table as shown here. After calculating the probabilities for X = 0, 1, 2, 3 and 4 we get the probability that X = 0.02, the probability that the random variable X takes a value of 1 is 0.07. The probability that X takes a value of 2 is 0.32, probability that X takes the value 3 is 0.43 and for X = 4 the probability 0.16.

Now this table here is called the probability distribution as it tells us the probability of all the possible values of X. So, now we will take these probability values and just draw them in a chart to aid our understanding of the probability values. On the x-axis here we have five random variable values and on the y axis we have our probability values. The chart here, where we plot are probability values versus the random variable values is called the probability distribution chart.

So, as you can see the probability distribution when displayed in the form of a bar chart look exactly like a original histogram just with a different scale. That is since probabilities can lie between 0 and 1 only, the values on the y axis are now restricted in the range of 0 and 1 .Each value here represents the probability of getting a certain number of red balls. So, this is what our probability distribution looks like in tabular form and in bar chart form.

As you can see clearly from this example a bar chart gives us a much better visual idea of what the probability will be hence this form is more commonly used. Now we can use this distribution which gives us how likely each value of X is and find the answer to our original question, that is would be in long run make money or lose money. So, if we recall from our office game setup we fixed our random variable X as the number of red balls we got in the experiment.

We then made a probability distribution for X which will help us to see how likely each of these values of X is to happen. Now as we said earlier, we want to use probability and find whether in the long term the game or a bet will be profitable to us the house or not, question is how will we do that let us see. Suppose thousand people are playing this game how much do you think an average person would win or lose in this game can you try and take a guess.

Say for these thousand people do you think you will be able to tell me how many people are expected to draw one red ball, how many people are expected to draw two red balls, three red balls, and four red balls. As you remember from before we know that the probability that X = 1 is the value 0.07. So, multiplying that by 1000 will give us the expected number of players who will draw one red ball thousand times 0.07 gives us 70.

So, in other words about 70 people will draw one red ball out of those four balls. Similarly, the probability that X =2 will be 0.32, so multiplying by thousand gives you 320. So, again, it means that 320 players on average would draw two red balls. Similarly for X = 0 thousand times 0.02 which gives us 20 people who will draw 0 and 12. 430 people will draw three red ball and 160 will draw four red balls. In other words, 160 people on average will win this game.

Question is so how many red balls are we expecting to see on average in our game? Remember 1000 people played this game and out of those 1000, 70 people had drawn one red ball. Similarly, 320 people drew two red balls, 0 balls were drawn for 20 people, three bowls by 430 people and four balls by 160 people. So, I can say that the total number of red balls we will get after thousand attempts at the game will be 0*20 + 1*70 + 2*320 + 3*430 + 4*160,the sum of all these events gives us 2640 red balls.

So, basically, we saw 2640 time a red ball in our thousand experiments. So, if we take the division of these two, we can say that we are seeing 2.64 red balls in one experiment on average. In other words, we can expect a person to draw 2.64 red balls in one game. And so, in nutshell the average value that we would expect to get for a random variable is called its expected value.

Let us continue our discussion of expected value.

Let us now formally define the expected value. Let us say that a random variable X can take values {x1, x2, x3 and so on till xn}. The expected value of this random variable X would be given by the expression X1*P(X=x1) + x2*P(X=x2) + x3*P(X=x3) + … so on, till … xn*P(X=xn). And remember probability of X = xi denotes the probability that the random variable X takes the value xi.

In our example the value that the random variable X can take {0, 1, 2, 3, 4} so the expected value of this random variable would be given by the expression EV = 0*P(X=0) + 1*P(X=1) + 2*P(X=2) + 3*P(X=3) + 4*P(X=4). We have already computed these probabilities before and plugging those values.

We get the expected value as : 0*(0.02) + 1*(0.07) + 2*(0.32) + 3*(0.43)+ 4*(0.16) = 2.64. So, again the expected value for random variable X that is the number of red balls is 2.64. However, I am sure you are wondering how can you get 2.64 balls in one attempt the number of balls would be 2 or 3 where here we are saying that on average, we expect to get 2.64 balls. Actually, it is not unusual to have an expected value that may never actually come out.

We say that the expected value is 2.64, it does not mean that we expect to get red balls in an attempt that many 2.64 balls. What is this value actually means, is that if you were to play a game infinite number of times then the average number of red balls you would get is 2.64 but, does this value really even help us with our question? We still cannot tell whether we would make money or lose in the long run. So, what should we have done differently to answer our primary question.

If you want to find out whether the house would make or lose money in the long run then we would need a different random variable probably a better random variable here would be the amount one after playing the game. Let us continue our discussion of expected value. So, let us define our random variable this way X equal to money 1 after playing the game once. This definition of X works well for us if we get a high expected value for it.

That would imply that on average players will win a lot of money in our game which is obviously not a good thing for us. On the other hand, a low expected value here would mean that on average the players do not win a lot of money. So, this new random variable X can take 2 values -10 and +150. This is because player wins 150 dollars for getting four red balls and loses ten dollars for getting any other outcome.

We know that the probability one X takes the value 150 is the probability of getting four red balls is = 0.16. On the other hand, the probability of losing the game when the random variable would take the value minus 10 is 0.84 which is the probability of getting any other possible combination of balls namely getting 0, 1, 2 or 3 red balls. So, we can go ahead and compute the probabilities of getting 0, 1, 2, 3 red balls, we could get the probability of getting 0, 1, 2, 3 red balls by adding this individual probabilities 0.02+0.07+0.32+0.43 = 0.84. In other words, the probability that a player will lose this game is 0.84. Applying the same definition of expected value as we saw before; we can compute the expected value for this event as (150*0.16) + (-10*0.84), which is
equal to +15.6 dollars. What this means is, that on average a player would expect to win 15.6 dollars by playing the game.

So, it turns out that the game is a disaster for the house if you want to make money you need to ensure that the expected value for the money won by a player is negative. This is to ensure that players lose money and the house wins in the long run. So, how can you do this? You can decrease the price money, increase the penalty or just decrease our player's chances of winning any of these choices will decrease the player's expected earnings from the game.

We can go ahead and fill in our rules and configuration of the game. So, you can now understand how casinos ensure that they remain profitable. The casino machines are all designed in such a way that the expected value of the money earned by the player is always negative. This means if we design the game such that the expected value of the money earned by a player is for example -10 or -20 dollars we can be fairly confident that the house will be profitable.

Let us look at one real life example of the expected value. We know that the expected value is the basis for insurance companies and we can do a simple example based on this concept. Based on

certain sources the company estimates that the probability that an accident will occur within the next year is 0.00071, on the basis of this information, what premium should the insurance company charge to break even on a 400000 dollars on a one-year term policy?

So, let us try to calculate the expected value that the company will need to pay to settle a policy. In case of an event of an accident the probability is 0.00071 and the amount to be paid is 400,000 dollars. However, in case of no accident the probability becomes 1 – 0.00071 which is equal to 0.99929 and in this case the amount to be paid is 0. Thus, the expected value will be X times probability of X when there is no accident plus X time probability of X when an accident occurs.

In case of no accident X times probability of X becomes 0 times 0.99929 which is 0 as shown in the table here. And in case of an accident X time probability of X becomes 0.00071 times 400,000 dollars which is approximately 284 dollars. These computations are provided in the table. Thus on average the company needs to pay about 284 dollars to settle a policy. Thus, the premium should be 284 dollars plus whatever return the company needs to cover such administrative overhead and profits.

Variance, we can calculate the expected value if we are given a random variable and its corresponding probability distribution. So, this answers our question about the central tendency of the distribution. However, you also learned that apart from the central tendency we also look at the spread or variability of the data. So, let us go back to some of our previous example where we created a probability distribution of the random variable X where X is the number of red balls in the bag.

We want to find out the variance of X.To calculate the variance we have a very simple formula for doing so. Basically, it is Var(x) = $(x - \mu)^2 * P(x)$ summed over all the values of X or alternatively $\sum_{i=1}^{n}(x-\mu)^2 * p(X = x_i)$ where μ or sometimes called as $\bar{X}$ or expected value of mean of the random variable.

Recall that for X = 0 the probability was 0.02, for X = 1 it was 0.07, for X = 2 it was 0.32 and for X = 3 it was 0.43 and for X = 4 it was 0.16. Here μ denotes the expected value that we had earlier

calculated so it will be the expected value of the number of red balls which we calculated to be 2.64 so, to calculate the variance of X it will be (0 - 2.64)^2 * 0.02 + (1 - 2.64)^2 * 0.07 + (2 - 2.64)^2 * 0.32 + (3- 2.64)^2 * 0.43 + (4 - 2.64)^2 * 0.16 which sums up to 0.8104.

Thus, the variance of this distribution is 0.8104 and the standard deviation will be the root of 0.8104 which would be 0.9002. Calculating the variance could be used to compare the spread of different probability distributions.

(Video Ends: 21:29)