

module_prob3_lecture13

Mon, 2/7 10:37AM 14:44

SUMMARY KEYWORDS

number, points, distribution, lambda, probability, x tilde, equal, factorial, binomial, cdf, receives, geographical area, crimes, probability mass function, similarly, average, day, occurrences, plot, calculate

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

So this point of the module level introducing you to some common discrete random variables, distributions. And then we've talked about the Bernoulli, which is success or failure in a single experiment. And then we've talked about binomial, which is the number of successes out of n experiments. So now we're going to talk about another common distribution, which is used to measure the number of occurrences of an event within a given timeframe, or a given area. So this is called the points of distribution. So the points of distribution was first proposed by a French mathematician called Simeon poisson. And what and the random variable here is the number of times that something an event occurs within a given time interval, or area. Right? So so say, you look at the day, and look at the number of times something happens during the day. Or you look at the geographical area, maybe a neighborhood or a city or a village, right? And you look at the number of times something happens in that neighborhood.



So, so x tilde here,



Sumon Majumdar 01:30

it can take zero, what are the possible values it can take? It can take 0, 1, 2, 3, 4, right? Because this is the number of times that the event occurs, and this just goes off up to because up to infinity, right, possibly infinity. So what we are really interested in is what's the probability that X tilde takes a particular value x . And let's look at that, right. So that's given by this expression here, that probability that X tilde takes a particular value x is equal to e to the power minus lambda times lambda to the power x divided by x factorial. And this is the probability mass function of the points of distribution. Now, various components here, we've already talked about x tilde, which is the factorial, we also saw that in the binomial, A , then x factorial is the multiplication of x times x minus one, and so on up to one. What is this number E . So this is the base of the natural logarithm. And it's approximately 2.71828. And this is also

sometimes referred to as the Oilers number. Now, see, this doesn't look right off the bat, something you dream up, right. But what people have found is that this fits a lot of situations quite naturally, one of the earliest places where this was fitted, was looking at the number of Prussian soldiers who were kicked by accident who were killed by accidental horse kicks. Similarly, people looked at the number of missiles that were fired by Germans into a particular neighborhood in London, and did this during World War Two. And they found that that fit this the points of distribution, and in fact, the poisson distribution has been used in a wide variety of contexts. So things like in the stock market, the number of times that the stock market goes up, over a given time period, let's say six months or one month, the number of crimes or murders in a geographical areas, think of a city. All of these have been modeled as points of distribution.



So to get

S

Sumon Majumdar 04:24

a better handle on this, right, so we'll do an example. But before we go there, I want to point out one thing. That's the main thing here. So see, we were interested in probability that X tilde is equal to x , right? And the main thing that determines this, right, there's, of course, this x , but the other important thing is λ . So but for the binomial, this was affected by two things, M and P . Few. Recall, right here, there's only one parameter. It's λ , and what is λ ? So λ is the mean number of times the event occurs within that given time. So if you're thinking, let's look at the number of crimes in a city, and the average number of crimes in a city over a year, let's say is 50, then we'll choose λ as 50. So let's do an example along this line. So let's look at on average, Pam receives five emails in an hour. And what we are asking is, what's the probability that you will receive six emails in the next time? See, this is not a typo, because on average, she receives five. So in the next hour, it's a number of English he receives is random. It could be zero, it could be one, two, could be 50. It could be 10. Anything, right? So what we are interested here is looking at what's the probability that the number of emails that she receives is exactly six. So this is a count variable. And we'll use the points of distribution to model this. And then the points of distribution, see, the average here is five, so we'll pick a λ is equal to five. So So remember, so the probability of X tilde equal to x , if it follows a points of distribution is given by this. So in this particular example, when we're looking at what's the probability that X tilde is equal to six, so we're going to make use of this formula, that's going to be e to the power of minus λ . So here, λ is five. So this is minus five. And this is λ , the second term is λ to the power x . So that's five to the power six divided by x factorial. So so divided by six. And this is a very complicated term to actually figure out by hand, because this he remember, is this number 2.71. Something something, right? So the best thing is to use a calculator. And if you do that, you will find that this is equal to point 146. Right? So the probability that she received six emails in the next hour is point 146. And you can ask similar questions, what's the probability that she receives knowing so this is the probability that X tilde is equal to zero, so all we have to do is take this formula and put an x equal to zero. Which is this. One thing to note here, see, zero tilde, it's taken to be equal to one. So all other numbers, remember, six, or zero factorial is taken to be one, remember, like six factorial is six times five times so on. Right? Then we traveled back up to one. So what zero factorial? So that's taken to be equal to one because we can't really do the same trick with zero because zero times what we're not going to be multiplying by minus one,

and so on, right? So this is taken to be one. Now if you calculate this, so this probability will turn out to be 0.06. Right? So this is a really tiny probability that she's going to receive no emails and the next time you're going to ask other related questions, what's the probability that she gets six emails or less. So what we're looking for really is the CDF at the point six. And what we're going to do is look at the probability of X equal to 0, equal to one and two, and so on up to six. And you can calculate this, this is a bit tedious because you have to calculate each of these right and then add them up. And if you do that, it's going to be point 762. And I'll show you in the next module, how you can do this using Excel. Excel gives you an easy way of computing for the cumulative or the probability mass function for binomial and four points. Now, what are other examples right. So you can use similarly things like suppose on average 10 crimes in a city per day, you can ask the so in this case, λ will be equal to 10. And you ask the question, what's the probability that there will be exactly eight crimes tomorrow? So what's the probability that this is equal to eight? Again, you can use the points of formula to figure that out. And this has been used in a lot of context, as I said, right? So things like number of website visitors in an hour, right? Because this is random, then some hours, there may be a lot of traffic, some hours very little. But if you want to know, what's the probability that at least 1000 or 1000 visitors visit the website, you can use the points of distribution to figure out that problem. So things like number of births in a day, number of bankruptcies in a month with all of these, because they're just so this is a car, these account variables of something. So these are all modeled, using the points of distribution. One other thing that I should mention before I move on from the points of distribution is that one important thing, and I'm not going to go too deep into the points of distributions derivation is that it's memoryless, in the sense that what happened, like let's say, you're looking at number of website visitors in an hour, right? The fact that very few, visited the website in the first 10 minutes, has no impact on the number who are going to visit in the next 10 minutes or the next 10 minutes or so. All right. So this is called memory less. Similarly, the number of facts, number of births in a day, right. So the fact that doing a lot of births in the morning of that day has no influence on the number of possible births in the afternoon of that day. So all of so suppose our distribution works when these events are memoryless. So what happens before doesn't have an impact on what's going to be happening in the future. couple of important things here. So I've already talked about the mean. So the mean of the points of distribution is λ . And interestingly, the variance of the points of distribution is also λ . So λ determines both its knee as well as its spread, which is the variance. Now, if you're interested in looking at how does the how do the probabilities look like? What's the shape of the PMF? Again, so in this case, it's going to depend on the single parameter λ , and you can plot it out. So Sophie plot out rate, I've done this for various values of λ . So the blue line is for λ equal to two, the orange one is for λ equal to five, and so on. And remember, λ is the mean number of four currencies, right? So when it's five, red, let's look at this here, right? So the highest probabilities are around five, it's at five and four, right. And before that, the probabilities are climbing up. And after that the probabilities are climbing, right. And by the time it reaches around 14, they have almost become zero. On the other hand, if you look at something like λ equal to 15, so this is λ equal to 15. So here, the probability of occurrences of the lower range, these are very small, the probability of occurrence in the range 25. And beyond these are very small, but in between five and 25, that probability is going up, it maxes around 15 1415 points, and then it comes down. And you can also plot the CDF and they get they tell you the same story. See, the CDF is very small up to five. So this is λ equal to 15. Right? And beyond that, it keeps climbing up but beyond about 25. The it's almost reached one. So it's this is almost surely all the occurrences are going to be 25 or so that's our points on distribution. And that's the last distributed common discrete value distribution that we are going to do

