

module_prob3_lecture12

Mon, 2/7 10:37AM 17:25

SUMMARY KEYWORDS

binomial distribution, probability, equal, x tilde, point, variance, values, vote, cdf, success, suppose, distribution, calculate, stock market, number, trials, cumulative probability, occur, question, characteristic trait

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

Since the last group, I introduced the idea of the binomial distribution. And in this case, we're going to talk a little bit more about it. Because the binomial distribution is a distribution that occurs very commonly. Anytime see, you're running an experiment, and you randomly pick 10 subjects, the question is, what's the probability that a certain number of them are left handed? Okay, or a certain number of them have a particular characteristic trait? In the stock market, suppose you pick a certain number of days, what's the probability that the stock market went up on exactly two of those days, or five of those days? So all of these are exactly in the realm of the binomial distribution. So let me begin by reminding you what exactly is the setup for the binomial distribution. So here you have n independent trials, right? independent in the sense that what happens in one doesn't influence what happens in the lab. Right? So when I was talking about the stock market example, so long as what happens on one day doesn't is independent of what happens on another day, then you're, that's the setup of the binomial distribution. Secondly, each has two possible outcomes, either the stock market went up or went down. Thirdly, all the trials have to be identical, they have the same probability of success, p . So that's the setup. And what are you interested in, you're interested in, in the variable, which is the total number of successes out of them, right, so that's your random variable X tilde. And what the binomial distribution gives you, is the probability that this x tilde takes a particular value small x . And that's given by n choose x p^x to the power x times one minus p to the power n minus x . In short, while I'll explain to you where these terms come from, I'm not going to entirely derive it. But I'm going to show you where this given intuition for from why the formula looks at as it does. So two things here. So the mean of the binomial distribution is given by n times p . And the variance is given by n times p times one minus p . Now these are easy to remember, if you recall the mean of the Bernoulli distribution for the Bernoulli. The mean was just p . It's the mean, just μ was just p , and the variance was p times one minus p . So the binomial distribution is this, multiplied by n . So it's the mean is n times p , the variance is n times p times one minus p . So the standard deviation is the square root of the variance. So that square root of n times p times one minus p now so these are the probabilities of each x , right? So you may be wondering how do these probabilities actually look? So we derive the particular probability in a particular example. But in general, how do these look? Right? So, so here is the graph, right? So these probabilities obviously depend on two things. One is n , the

number of trials, and secondly, P , which is the probability of success in any particular trial. So the shape depends on n and p . So these are the two parameters of this distribution. This as you change an MP , the look of this distribution changes. So what I've done here is I've plotted out for different values of NFP . So the first one, this light blue one, this is for n equal to two and p equal 2.2. So this will run from zero to 10. Right so here the possible outcomes are either zero successes, one success or up to 10. But what do you see is that most of it like centers around these early parts between zero and five, right? So that's, that's the probability is highest at this point, and then it grew rapidly comes down. And this is around the point six or seven, it really becomes almost become zero. What about the next one, this is n equal to 30 and P equal to point two, and that the following one is the same n equal to 30. But the probability of success here is point five. So see the orange one, right? So the higher probabilities occur early on. And then it slowly dies down and becomes zero around here. Right? While when the probability goes up, 2.5. Right. It has low probabilities early on, but it has middling it has the probabilities Keep Going up Going up Going up around the this is on the 15. Mark, right. And then again, it starts coming down and slowly dies out when it's about 25, or so. So Plaga, two more, this is an equal to 50, which is an equal 2.5. And this is the last one, this one here is n equal to 50, and p equal to point eight, right? So when p is equal to point A, when the probability of success is high, so what do you really expect is more successes, right? So you see that the probability because each of these are given the probability of the corresponding value, right? So here, you see that more of so the probability of about 40 is the highest in this case, right? And then comes the probability of 41 42 43, and so on, and then it's slowly dying down. So the probability of getting a value like 49, or 50, this is very, very small. Same thing. Now, if we look at the yellow one, which is n equal to 50, and p equal to point five, see its values like 25, which have the highest probability of occurrence, right? And by the time you come to about 40, or 45, and so on, right, the probability has really become close to zero. Now you can also ask the question, how does the cdf of this looks like remember, the CDF is the probability that X tilde takes values less than or equal to x . And the way to calculate it is that you add up the probabilities of all the values which precede x , right, so each of these are values of. So if you look at y , starting from 012, so on up to x , right, and you add up all of those probabilities, you get the CDF at the pointers. So this is going to be a bit cumbersome, because then you have to calculate the probability at every point and then add these up. Luckily, you can do this in Excel. And in the next module, I'm going to show that to you. But right now, I've generated a graph using from Excel using those values. And see. So each of these correspond to a different set of parameters. This one corresponds to n equal to 50, point D. So if you look at this, right, this is the cumulative probability this is very low, this is almost close to zero, up to about the point 33. And then it starts rising, rising, rising, and the cumulative probability becomes almost equal to one by the time you're reaching about 4445. On the other hand, if you look at this the same n equal to 50, that p is lower, it's equal 2.5. In this case, it starts rising a lot earlier, it starts rising around the 1617 mark, and then it flattens out reaches almost one by the time 35 Comes right. So these depend on n and p the probability of success now, let me do one more more example of the binomial distribution. So since the recently concluded the Canadian federal election, the voter turnout was about 60%. Now suppose you pick eight adults at random, and ask the question, what's the probability that exactly for the sample? So this is exactly the the setup of the binomial distribution. So here, you have eight trials, right? So n equal to eight. What's the probability that any one of them voted is point six? B equal 2.6? And what are we asking the question that probability that X tilde is exactly four. So this is exactly my binomial with n equal to eight and p equal to point six. And I can calculate that using my formula says a choose four P , which is point six to the power four, and one minus b , which is point four to the power four. So first thing if I'm calculating 8 choose four, that's eight factorial over four factorial. So eight factorial is eight times seven times 6, 5, 4, 3, 2, 1. Write four factorial is 4, 3, 2, 1. And again, the second four factorial is again 4, 3, 2, 1. Now see, this cancels with this. Three times two,

that's cancels with this four cancels this. Right? So this turns out to be seven. And if you calculate this all out, this will turn out to the point. So what that means is that in this federal election, right, if you picked eight people at random, what's the probability that exactly four in the sample voter is 23%. So now, let me try and explain to you a little bit how this binomial distribution comes around. We'll see if we're looking at so in this right, so if we want to look at for voted, and for dental. Right, so the probability of these guys, right, each of them are point six, right? So if the probability that all four voted is point six to the power four, the probability that the four didn't vote, each of them that didn't vote is one minus point six, and also probability that all four didn't vote. So that would be one minus point six to the power four. Right? So this is the probability of four voting and four didn't vote. Right? So that explains this term here. And this, right, so what's this? What this comes from is the fact that if you're picking eight people at random, and you ask the question, what's the probability that exactly four in the sample fold? Now, it could occur in many ways, right? It could be that the first four voted the roommate last four did. Or it could be that the first one voted, the second one didn't. And right, or it could be that the first two of order, the second two didn't like next student vote, these two voted, right? So there are lots and lots of possibilities that so how many such possibilities are there exactly, is given exactly by this age useful. So this is where this term comes from. So I'm not getting into what this is in the realm of what's called permutations and combinations. So I'm not going to get into that in the course much, right, but what I'm trying to give you an intuition for is where this formula comes from. Now, in this particular example, you could then ask the question, what's the probability that for like, here, I ask the question, what's the probability that exactly four vote? I could ask the question, what is four or fewer vote, right? So then you add up the probability that X tilde is equal to 0, X tilde is equal to one, and so on up to four. But which would be the cumulative distribution? Again, one can it's going to be a bit tedious, but one can do. What else? You can ask, what's the mean number of voted. So here n is a p is point six. So that's four. Similarly, what's the variance, it's going to be n times p times one minus p , so that A times point six times four. That's going to be one point. So hopefully, this has given you a good introduction and a good idea of the binomial distribution. And because as I said, it's used a lot in interviews. So let me end this clip. with the, with the clicker question. So please, you may wish to stop the video at this point, and so hope you've got a chance to do it. Right. So what it asks is that suppose the probability that any given email is spam is point two five, so here p is equal to point two five. And suppose you receive 80 emails in a day equal to 80. What is the mean number of spam emails that you will receive? Right, so the mean so this is a binomial distribution, we have 80 trials out of each each of them being a spam is point two five. So the mean number is n times p . So that will be 80 times point two five which will be so the correct answer. So let me stop the clip here. And the next clip, I'll be introducing you to another common distribution, which is called the poison distribution.