module2_lecture3

Mon, 12/27 3:51PM 15:18

SUMMARY KEYWORDS

summation, points, summation notation, data, notation, formula, average height, equal, clicker question, trees, divide, sum, height, denotes, wealth, squared, variance, intersection, family, number

SPEAKERS

Sumon Majumdar

Okay, so so so far, what we have done is I've introduced you to the summation notation. And the idea is very simple. That a notation like this, where we are summing up the sigma over Zi, from 1 to N, this just means the sum of Zi's running from i equal to 1 to i equal N. And as I've told you that it need not start from 1, right, you could sum up from 5, i equal to 5 through N. So what that means is, you're adding up from Z5, Z6, so on up to ZN. And this sort of things come up in different applications. And I've already to like in the last clip, you saw examples from number of COVID cases of a particular month, so either during the summer, or during the second half of the year. It can come up in other contexts as well, where you only add up certain section of individuals. So for example, here is an example from a village with 40 families. And these are, let's say, ranked in terms of their wealth, starting with W1, this is family 1's wealth, this is family 2's wealth, family 3's wealth, so on up to family 40. And I've ranked them that, that family 1 is the poorest, they have the least wealth, then comes family 2, and so on. And family 40 is the richest, they have the most wealth.

So what if we wanted to ask ask the question, so what's the total wealth held by the top 10 families? So in this case, what we would like to add up is the wealth held by the top 10 families, right. So we want to add up the wealth, but we don't want to add all the family. So which are the top 10? So it's the families which are 31,32, 33, so on up to 40, right. So these are the, the richest 10 families. So if I want to just find out the total wealth held by these families, I will add up from i starting from 31, and going up to 40. So this is this is the idea again of the summation notation. Now just to be, again, to be absolutely clear that we are all on top of this notation. So what I have for you is another clicker question. It's a pretty simple one. But what I would like you to do is stop the video at this stage. And go over the clicker question, just because we're going to get into a little bit more involved applications of the summation notation. So please stop the video at this point and try the clicker question.

So I hope you got a chance to do the clicker question. So what it asks is, is this set up, that there 24 hours in a day, and Ci denotes the number of cars which are passing through in through an intersection in hour i. and this type of studies are often held for planning purposes, whether you want

to put up a streetlight, whether you want to put up a stop sign in a particular intersection or not. And what this question just asks you is that which of the following expressions denotes the total number of cars that pass through the intersection during the day? So the day has 24 hours, so if you want to add, if you want to find the total number of cars that pass through this intersection during the day, you add up C1, that's the number of cars in hour 1, then in hour 2, so on up to hour 24. So this is the summation of Ci, i running from 1 through 24. So the right answer to this particular clicker question is C where it adds up the Ci's starting from i equal to 1 to 24. So now that we have got a handle about the summation notation, let's use it in a couple of applications, which are slightly more involved, but things which you will see in a lot of other contexts.

So one of these is this that suppose a forest has 200 trees, what is the average height of these trees? Now the average height to find the average height, all you have to do is sum up all the trees heights and divided by the number of trees that there are, which is 200. So how would we do that in, using our notation, right. So if we have to sum up every tree is height, right, so if Hi is the height of tree i, okay, and we want to find the total of all the trees heights, right, then we want to add up starting from i equal to 1 to i equal to 200. And then this gives the total height of all the trees and then just divide this by 200. So this is going to be the average height. So I've used my math, my summation notation to denote this in a pretty succinct way.

So, so again, just to reiterate that the total height is the sum of every trees height. So this is the summation of Hi, from 1 to 200. And then the average height is, is this divided by 200. And you can write this as 1 over 200, right, so you can bring this thing in forward. So that's 1 over 200 times the summation of the H's. So that's the average height in a forest with 200 trees. Now if we generalize this example, and ask what would be the average height in a forest with N trees, you do exactly the same. You would add up the height of all the N trees in that forest. That's that particular summation notation. And then divide the whole thing by N. So that would be 1 over N summation Hi.

Now, so that's a pretty simple formula for finding the average height of trees in a forest with N trees. Now, you can generalize this, and in statistics, right, they may not be trees, it could be N data points, right, which could be data about anything. It could be the data of the heights of N people in a city. It could be the incomes of N people in a city. It could be the number of COVID cases in N cities across Canada. It could be the number of crimes in N cities across Canada. Okay. So in statistics, these are called data points. Right? So this is, let's say the point for city 1, this is for city 2, this is for city N. Or this is the height for person 1, this is the height for person 2, it's the height for person N. Right, so depending on what data you're talking about, these are data points. And how do you find the mean of N data points? It's again, the same thing that we did with the trees, we add up all of them, right, and then divided, divide that sum by the number of data points there are. And now we can, given our summation notation, we can write this pretty succinctly, because the summation, this X1 through XN, this can be written as summation Xi, index from running from 1 to N, right? And divide this by N. Divide this whole sum by N. So that's the formula for the mean of N data points. This is in statistics, this is given by X bar. So we're going to do statistics in the second course, but I'm just giving you a preview, that just knowing your summation notation, you can already see how the formula for the mean of N data points, which is going to be an important aspect, is going to look like.

And so, so again, reiterating here, the mean of N data points X1 through XN can just be written as the

summation of the Xi's from one to N times 1 over N. So this is a standard formula, I'm sure you're going to encounter it at some point in your university. And all it means is, is exactly what we just discussed, this summation is just X one plus so on up to XN, right? It's the summation of these N data points divided by N.

Now, another thing, which you're probably going to also encounter is, so this is the mean, another thing that you're probably going to encounter is also the variance. Variance tells you how dispersed are data points. And again, you're going to cover it in much more detail in the second course, where statistics is going to be a big part of the course. But right now, let me just introduce you to the formula for variance and talk about what it means because it's a summation form. So the variance from N data points is given by this formula, which is summation of Xi minus X bar, whole thing squared, right, and the index running from 1 through N, and the whole thing multiplied by 1 over N. Now, at first glance, you may be thrown off a little bit by this formula looking considerably more complicated than the things that we have done so far. But if you go slowly, it's exactly the same notation that we have used so far. So let's, let's decipher this. So what does this mean? Right? So let me keep the 1 by N outside, right, let me open up a bracket. So the index when Xi is equal to 1, right? So that means, let me put that in. So that means that is X1 minus X bar, whole thing squared, plus, now when i is equal to 2, it's X2 minus X bar, whole thing squared. And this is how it goes up to the final one, which is when index i is equal to N. So this means XN minus X bar squared. So what it means is that say, if you have your data points, suppose your data points are, are something like this. Okay? Let's say they are 10. I'm just making up some data points of 8, 4, 40. So what this tells you is that firstly, you compute the mean, right? So the, to get the mean, you add all of these up, divided by the number of data points, which is 5, right? Whatever the value is, that's your mean.

And now, when you're going to compute the variance, all you have to do is let's, we have to use this particular formula. But it says, okay, there are 5 data points here, so this is 1 over 5. Okay, then the first one says X1 minus X bar, so my X1 is 10. So that will be 10 minus whatever is the value of X bar that I have got from here, right? And then I square it. Then I go to my second data point X2, which in this case is 15, right? And I subtract X bar, and then I square it, and so on. And I get a value by doing this computation. But the main point is that here is the formula for variance, and it is a summation form. And all it means is that okay, for just patiently put in the index every time you see an i, and then go on to the next one, then go on to the next one, and see where the index runs up to. That gives you what that summation is. So what I've tried to show you in this particular clip is that applications of the summation formula crop up in lots of different circumstances and we'll see more examples in the next clip.