

module_stat4_lecture3

■ Thu, 2/17 4:18PM ⌚ 11:41

SUMMARY KEYWORDS

categorical variable, population, daily interactions, quantitative variables, sample, estimate, variable, parameter, fraction, average, person, question, number, interested, tax increase, commuting, suppose, examples, greek letter, statistic

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

Now suppose you have gathered together a representative sample group of people, they have decided on that you're going to sample or a lot in the forest that you're going to look at trees there. So, then comes the question that what exactly the way to do it? And this relates to the basic question that what are you trying to find about the population? So let me introduce you to some terminology in this context, because this is very important, because what is it exactly that you're trying to find out about the population? That's going to determine what sort of question that you're going to ask yourself? Or what sort of measurements are you going to do for the trees that you want to sample in this forest? So first, is this what's called the parameter μ , right. So this is some characteristic about the population that you're trying to find out about. So this is called the parameter. And this is usually denoted by a Greek letter. The, the most common Greek letter that's used is μ . So what sort of characteristic are we talking about? So it could be, for example, the amount of debt that an average Canadian, or it could be the fraction who support a tax increase, or the fraction of Canadians who drive over the speed limit. Or it could be something like remember, in the very, very first module that we did for one, we tried to build up a model of infections, right. And in that one of the most important parameters is the number of daily interactions that a person of average person has, because the more the interaction, the faster is the spread of the virus. Right. So one thing you may be interested in is what's the number of daily interactions on average that a person has. So this, all of these would be examples of a population parameter that you're interested in finding. Now, one other thing I want to point out while I'm talking about this examples, so see, this is the number of daily interactions, this is the amount of loans right. So these are all examples of quantitative variables. So earlier, we talked about types of variables. So these are quantitative variables. Now fractional support a tax increase fraction who drive over the speed limit. So these are categorical variables, because would you would ask a person is Do you support a tax increase? Yes, or No? Right? So these are examples of categorical variables. So it's, the fraction itself is a number, right? But the underlying basis by which you're going to compute that fraction? That is a categorical variable, because people will say, Yes, I do sometimes drive the speed limit, or No, I never. Now, once you've determined what you're trying to find out about the population, right, you try and measure this variable in the sample, right, this or a related variable? The reason why I ask is why I say this is the following. You see if I number of daily

interactions, right, so this is a non controversial variable. So people will probably readily answer your question about how many daily interactions Do you have? But suppose you ask purse, a person? What's the amount of savings that you have in your bank account? Or how much did you earn this week? So these are more touchy questions, and not everyone may be comfortable answering that letter. Sometimes. In the real world, what you do is you ask a related question, that would you say your income last week was between 500 to 1000? Or was it between 1000 to 2000 2000 to 5000, or so on? So, when people are actually doing surveys, they oftentimes like if you if you fear that the variable x is something that people may not be willing to answer directly, you ask something related. And then once you have this variable, what do you do that So use this data from the sample to form an estimate of the population. It's maybe it could be the mean of the answers that people say about how much loan they have. Or it could be the fraction of people in your sample who say that they support a tax increase. So these things that you calculate from the data, that's called a sample statistic. So remember these two terms, because we're going to make use of it quite heavily in this module, right? So when you're, when you're looking at the population, what you're trying to find out, is this parameter μ . But obviously, you cannot examine the whole population, so you just examine a small part of it, which is the sample. And once you collect data from the sample, and estimate something, that's called a statistic. So an example of it could be the mean. So, so for the population, it's the term is parameter and for the sample, that is statistic. Now, okay, so suppose you have the sample, you measure the variable x , right? Now what? So you have a, you, you, you collected the sample of size n , right? You measured, you ask them questions, right? So that was the variable x . Now, what do you do? So, now, it really depends on what is that variable x that you're trying to, you're interested in. So, if it is a quantitative variable, numerical variable, then the most usual thing that people look at is the average of that value in the sub. So suppose it's a , it's an example, the amount of household debt, okay, or the time spent commuting? The, the number of friends one has, right, so these are all examples of quantitative variables. So in this case, you take the average of the sample, right? And this is an estimate of the mean value of x in the population. So if you're interested in is what's the, the average number of friends for Canadian are often American, or the average number of friends that a university student in York University has? Right? Or the amount of time that that an average student in York University spends commuting to school? So this $\bar{x}$ is an estimate of what you're trying to find for that population? Well, similarly, if x instead is a categorical variable, right, so think of a categorical variable, like people who smoke, right, or those who support a particular policy P , people who are single, right, so these are all categorical variables, because what you have is either someone is single or not, right? Either someone smokes, or doesn't smoke, someone supports the policy P or doesn't support it. So in this case, would you look at in the sample is the proportion of people in the sample with x equal to one? So the proportion of people in your sample who are single, right, so that is an estimate of the mean. Now mean? This should not be the mean number, right? This is an estimate of the fraction of people with x equal to one in the population. So the so this part is very simple, right? If you have a quantitative variable, you look at the mean, if it's categorical variable, you look at the proportion of people in your sample, the fraction of people in your sample who have that particular characteristic. Now, of course, you may be wondering, then, this seems very simple, right? I look at my sample, see, what's the fraction who are smokers? Let's say that's 20%. And then I say, oh, that's an estimate of what's in the population. So 20% of the population, therefore must be smokers. Right? So that's my estimate. Or if I, if I look at my sample, and I see, okay, on in the sample, the average time that a person spends commuting daily, maybe an hour and 10 minutes. So then my estimate of how long people in the population, right, and the population that were interested in studying, spent commuting, my estimate is an hour and 10 minutes, right. So this seems very simple, right? Why do we even need statistics for this? So what I'm going to show you in the next clip, is that statistics tells you why this is a good

estimator. And not only just tells you this, but it also gives you an idea of how much confidence should you have in that estimate. So suppose your sample 1000 people, and you've got an estimate that 20% of them smoke, and use that to form an estimate for the entire population, that entire population 20% of the people smoke, how much confidence should you have that this number that you found, is a good number or not? So let's stop this clip here. And in the next one. I'm going to talk about issues issues like this. And the basis of that will be from probability