

module_stat4_lecture2

■ Thu, 2/17 4:18PM ⌚ 14:17

SUMMARY KEYWORDS

sample, population, bias, entire population, sampling, statistical inference, small fraction, characteristics, landlines, survey, canadian, issue, soup, important, students, similarly, university, representative, banquet, entire

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

Let me reiterate the main idea of statistical inference that there is, for some reason, it's impossible or impractical to examine everyone that we are interested in knowing about. So that's called the population. So, in that case, what we need to do is we need to survey a small segment, a small fraction of the population, and from that segment, make an inference about the entire population. So that's the main objective of statistical inference. Now, before we get there, so the first question is, what is the population that we are talking about? So so far, I've been talking about human population, for example, the population of Canada. So if you want to know what's the average wealth, the wealth, average wealth of a Canadian, you need, your population, then would be the entire population of Canada. But it may not be the entire population. So for all of Canada, for example, if I'm interested in knowing how well the students in my class, think about my teaching style, in that case, the population would be all the students in my class. But as they say it like in a survey, not all of them respond, so a small fraction of them respond. So from that small fraction, I have to figure out what the entire class thinks. Now, the population needn't even be humans. So for example, if you're in geography, you may be interested in knowing about, about the health of trees in a particular forest. In that case, your population is all the trees in that forest. But obviously, you cannot go around counting and measuring every tree in the forest. So you need to look at a sample. Or you may be interested in knowing the water quality in all the Great Lakes. But obviously, you cannot look at water quality from everywhere. So you take samples. So this issue of sampling, it comes up, you know, nearly all areas of the social sciences. So sample is a small fraction of the entire population. Now, when you're drawing a sample, right, so it's so what are the main issues that arise? So first issue is the issue of what's called bias. And so for example, if I want to know how well the students think about my teaching style, and I only asked the students who got an A in the course, obviously, what I'll get is a very biased perspective, I'll get the perspective of the students who did well in the course. Similarly, if I only asked the students who fail the course, then I'll again get a biased perspective. The same issue arises in other cases, as well, right? So. So it's important that when you're drawing a sample, right, to eliminate as much bias as you can. So what what sort of biases, right, the bias may not always be accurate. So for example, is say you pick a lot of surveys these days are conducted over the phone. So suppose you look at the phone book and pick out numbers from there and call people. And you think, Okay, I'm just

picking up people are random and picking them up, right? So they may not be a bias. The problem is that a lot of people these days don't have landlines and landlines are the numbers that's on the phone book. So especially younger people, many of them don't have landlines, they only have a cell phone. Right. In this case, if you're only depending on the phone book, you may be missing out a lot of the young people. On the other hand, if you do your survey by the internet, you may be missing out on the older people who don't use Facebook maybe as much as the younger people. So this so people will actually run the service. They have to take these sort of biases in mind. And what the what they try to do is that one solution to get around some of the biases that may not this telephone versus internet is that you randomize. That means, for example, if you if you want to know about, about the time usage on a cell phone, right, you don't just ask your friends, because you know that your friends are probably like you, if you want to get a perspective about the entire population, you want to look beyond that. And you want to pick people randomly, and without any particular pattern, so that you get, you don't you kind of average out, average out all the facts, and then the sample starts looking more and more like the population. And this leads me to my second issue of sampling is that you need to make sure that the sample is representative of the population. Again, this is related to bias, right. But it's, it's a little bit more than bias. So what it means is that suppose you want to look at the, the wealth of the average Canadian, or the income of the average Canadian, right? So you need to make sure that you're sampling both younger people, middle aged people, older people, right, because typically younger people, they starting out in their careers, they probably have lower incomes than people in the middle aged, right, who are more experienced, and therefore their salaries are likely to be higher. Similarly, you need to look at men and women, you need to look at immigrants, non immigrants, people of different racial backgrounds, different ethnicities, right? If you want to get a picture of the entire population, so you need to make sure that the sample is representative of the entire population. So and to do that, so the way to do that is that you try and match many characteristics of the population. So what do I mean by that? Right? So you want to suppose you want to do a sample of 1000 people? Right? And you want to, you want to make sure that it's representative of, of the population that you have, right? So what do you want to maybe do is first start out with if you want to do it, ensure that it's representative of the gender, maybe you want to pick 500 males and 500 females. And among the males, you can pick randomly among the females, you can pick randomly, but at least it's going to make sure that you're on the gender issue is going to look like the overall Canadian population. Same thing, if you want to do it now by age, right, you want to maybe slice it up and make sure that a certain fraction or a young certain fraction or middle age and certain fraction are old. Similarly, you may want to do it by education that makes sure that a certain fragment of your populace of your sample has university education, some which are below University, some only have school, and so on. But the problem is that as you do more and more and more characteristics, your brain you're cutting up the sample into many, many, many different ways. And soon it you may have difficulty slicing it anymore. And the issue again related to bias that you may be controlling for all the characteristics that come to mind, but there may be other things which don't even occur to you, and which may bias yourself. So professional statisticians try and keep as many of these characteristics in mind as they can. Now the third thing is that how large of this random sample do we need to choose? Do you is this a big sample? Is this a big sample? How big a size of the sample do you use. So in general, in general, the bigger the size, the size of your of your sample, right? So better is going to be the confidence in the estimate of the result that you find. So let me give you an example from Got this example in the textbook and I find it very act. So suppose you're making a soup or some kind of a dish. And you're worried about the salt or some taste of that soup. Obviously, you cannot eat the entire soup because you're getting ready for the guests. So typically what you do is you take a spoon and take a sample, and use that to measure the amount of salt or the tastiness of that soup. Now, suppose you want to you making that same

soup not just for four people, but for 40 people for a big banquet? So, so the question then is that do you? Do you really like if you want to get a sense of the taste of this bigger pot of soup? Right? Do you need really need a bigger ladle to to to taste to take the sample? Not really, right, what do you need is you can take that same small spoon, but you need to make sure that you you taste take the sample from various parts of the pot. So So coming back to a her real sampling, what matters is the absolute size of the sample. Right? Not its size relative to the population. So what I mean is that if you have a sample of 200. So this is going to give you enough information about if your population is your university, right? The amount of information that this 200 students is going to give you about your university, it's going to be the same amount of information as a sample of size 200 gives you about the entire Canadian population. Right? So what matters is the absolute size of the sample. Not that its size relative to the population, the fact that 200 is small compared to the entire Canadian population. Whereas 200 may not is okay for a university population, that doesn't matter. What matters is this absolute size. So obviously, a sample of 300 is going to give you better information, better quality information as compared to 200. But that's true of whether you're looking at only your university, whether you're looking at only your neighborhood, or if you're looking at entire county. But what does matter when you're looking at entire Canada, or like a larger population, right? So the fact is that a larger population is also more diverse. So what then it becomes even more important that a sample is representative of all right. So again, like when you're making the soup pot for the entire banquet, it's important to take samples from different parts of the red, the bigger spoon doesn't help. But what helps is taking samples from different parts of that population. So these are the main issues with sampling. Obviously, sampling is a big subject and people professionals to decision spend a lot of time getting the design of the sample correct before they actually go out into the field. Right. But what I've touched on is three very crucial and important issues in sampling. One is to deal with bias. Secondly, make sure that it's a representative sample. And thirdly, it's the absolute size of the sample, not how big it is relative to the population that matters.