

module_stat4_lecture1

📅 Thu, 2/17 4:18PM ⌚ 13:32

SUMMARY KEYWORDS

surveys, statistical inference, data, population, canadian, statistics, sample, entire population, organization, module, inference, theorems, conduct, opinion polls, canada, impact, questions, small sample, government, date

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

So welcome to a new module on statistics, this will be our last module on statistics. And it will bring together several of the elements that we covered in the previous modules. So what is this module about? It's about statistical inference. So the basic idea is this that so far we have looked at data, it's a data on 100 people, 1000 people, 20, players, or so on. And then we have looked at how to visually represent those data using graphs, looking at Mean, Median variance and things like that. But oftentimes, and most of the times in statistics, we're not just interested in knowing about those particular 100 people, right? We're not interested in the 50 people that you polled about whether they like the performance of the current government or not, what you're really interested in is knowing about what the entire population of the country is thinking about in terms of the performance of the government, or of organs or on some other variable? So the idea is that how can you take data that you collected for a small sample for a small group of people and use that to make some inference about the population at large. So that's the idea of statistical inference. So this is a really big area. And it's an area that is also technically quite demanding in lots of different ways. So I'm going to cover only a small fraction of it, and kind of introduce you to the main ideas of statistical inference. So let me talk about the outline of the module. So we'll start with the basic introduction to the topic, what is the main questions there? And how the basic idea behind statistical inference, then I'll spend a little bit of time about sampling, because a lot of the issues will surround how to take data from the sample and use that to infer about the population. And there are two main theorems which underlie this main theory of statistical inference. And these two theorems come from proven from probability. And so we'll connect back with the module that we did on probability earlier. So the two main theorems are the law of large numbers and the central limit theorem. And then I'll be talking about the main main heart of statistical inference, one of which is confidence intervals, the other is hypothesis test. So let me begin with again, reviewing something that we talked about earlier, that what, what does statistics do in the social sciences? So we use statistics to answer what are called what is questions, things like what is the average age of the Canadian population? Or what fraction of the Canadian population smoke, right, maybe drink, or own a smartphone, sample support a particular political party or support a particular policy, say, a tax increase to fund free childcare, or a vaccine mandate? Or, or things like that? So one of the rules of statistics is to answer the sort of what is questions? The other rule of statistics is

also to answer what's called policy impact questions. So for example, what would be the impact of subsidized child care on children and parents? So this is an issue which is being debated a lot in Canada these days, that if you provide subsidized childcare, what is the impact on that on the children's development? And on the parents employment? Similarly, what is the impact of providing free laptops on children's education? So different people may have different beliefs about this policies? So it's the role of data to actually answer these questions in a more meaningful quantitative way. So that's the role of statistics in the social sciences. Now, if you want to answer these sort of questions, right, so What's the average age of the Canadian population? Or if you want to ask the question is the average age in Canada higher than in Japan? So you need data? And to answer this question, you need the data for the whole population. So if you want to know, the age of the Canadian population, or the number of kids that an average Canadian family has, or, or things like that, right, so you need data about the entire population. The problem is that that sort of data is relatively rare. So for example, if you want to know, what fraction of the Canadian population supports a particular political party, you wouldn't go out and ask every possible Canadian, right? So for that would require a tremendous amount of effort, tremendous amount of time, a lot of resources, right? So you don't. We don't do that. So data for the entire population is relatively rare. So it's not that it's not, it's not there at all. It's so for example, classic cases, the census. So for most countries, a census of the entire population is conducted every 10 years. In Canada, it's every five years. But that gives you an idea of the population in that particularly. And then you use like things like birth records and death records, to extrapolate and find the population in the in between years. Similarly, for example, if you are looking at trade data, right, exports and imports, all of these have to go through customs. So customs have these records of exports and imports. And from that you can get trade data. Now some of this data could be confidential, some of it could be available. And it also depends on what is that you're interested in study. So for example, if you want to know the number of points that were scored in an average game in the NBA, or in the NHL, in that case, the population is all the games that were played in the NHL or NBA that particular season. So that sort of data from the sports leagues, that's usually easily available. On the other hand, if you want to know, what's the average income of a Canadian person, right, so technically, the tax authority should have that data. The problem is that that data is probably confidential and you cannot access. So most of the time, what do you do is, and the most common way is that instead of the entire population, you draw a small sample. Right? You take a fraction of the population. So for example, if you want to know, how many how many minutes does an average person spend on his or her cell phone? What do you may do is just ask a few of your friends. And from that try to form an estimate about the rest of the population. This is what most opinion polls surveys do exactly right. So this sample, or ask questions of a small set of people and use the information that they gathered there to figure out about the entire population. So who conducts the sort of sample surveys or polls? So the common ones are government organization. So for example, in Canada, Statistics Canada or stat can do conduct a lot of surveys. I looked at their websites, they have about 180 surveys going on. Right? In the US. It's the US Census Bureau in India, it's the national sample survey organization. In China. It's the National Bureau of Statistics of China. So most of these government organizations, they conduct a lot of surveys on a lot of different areas. So for example, if we look at Statscan, they look at do the labor force survey. That's the LFS they do the CSS, this is the Canadian social survey, then they have a national transport survey, and things like that. So there are similar community survey in the US. And there are lots of different surveys right now going on about the impact of COVID and the pandemic. So these are common surveys that are conducted regularly by government organizations. Who else? Who else is things like this the Pew Research Center in the US, I've talked about them in an earlier module as well. They do a lot of opinion polls, it said they do it on a diverse set of areas, like immigration work, the economy, religion, they even ask questions like, what's the meaning of life. So they really do a diverse set of

surveys and go to the website, you will find discussions on a lot of them. Nielsen is a commercial organization, which does surveys on TV viewing on the online usage and things like that. Gallup does a lot of surveys for firms, and so on. Right. So these commercial organizations are, are another group of people who do sample service. One other source is individuals. So for example, most courses that you do that you take, have a survey at the end of it. Right? Many companies, they have a customer service survey, customer experience survey. So these are also surveys, and these are also generating data. Not everyone takes those surveys. So for example, when I teach a class of 100 people, and at the end, there is a survey about the class. So typically only 20, or 30 of the students take. But based on that, I try and make inference about what the rest of the class is thinking. Did they benefit from the course? Did I go too fast? Did I go too slow? How should I adjust my teaching style? So the main heart of statistical inference? Is this two questions that how do we use data from the sample to make inference about the entire population? And secondly, how much confidence do we have in our findings? So this this inference that we've did right, is it just a guess? Or is it more than a guess? Right? Are we doing better than, are we doing better than just the groundhog? Or Paul the Octopus or? Or some other random guess? Right? So statistical inference is about these two questions, right? What is how do we use data from a sample to learn about the entire population? And secondly, how much confidence do we have in that fight?