

module_stat1_lecture4

Fri, 1/21 12:23PM 12:11

SUMMARY KEYWORDS

data, vaccinated, population, cases, question, column chart, categories, sample, called, entire population, unvaccinated, census, qualitative, subset, status, inference, fully vaccinated, conducted, pictures, study

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

So so far I've talked about data, variables and the different types of variables. Now I'll start this clip by talking about a very important distinction in statistics between what's called population and what's a sample. So what you'll hear these terms quite a bit. And let me clear that distinction between the two, in the context of the type of data that you have, and the type of data, the type of things that you want to do with that data. So what's a population? So the idea of a population is the entire unit of study. So for example, you may be interested in that some facts about the entire Canadian population, or about the entire population in North America. So the entire unit of study that's called the population. Now, what's a sample? So sample is a small segment of that population. So here, this is my population. And if I look at a small segment of it, or a subset of that population, that's my sample. So let me show you were like, actual examples of population and sample. So see, if you remember I talked about the Canadian population. So how do you get data about the entire population? So the entire unit of study, so a common example of that is the census. So the census in Canada is conducted every five years, and it tries to enumerate every man, woman and child in Canada. Now, it may miss some, because you know, someone may not respond to the questions. For some reason, you cannot find someone, but more or less, they count everybody in Canada. But of course, the census is a, it's expensive, it, quite requires a lot of effort. And therefore it's conducted every five years. In some countries, for example, in the US, or in India, it's conducted every 10 years. But that gives you a picture of the entire population. Similarly, the COVID case numbers in Ontario that I talked about, gave an example in a previous clip, that's the COVID case number for entire Ontario. So that's population data. Similarly, if you're, say, interested in studying basketball games, and you have the data for the entire NBA season. So that's the that's your population, because you have the data for the entire unit that you want to study. But these examples are relatively rare, because, for example, suppose you're interested in a question that what fraction of Canadians smoke that so that's not a question that is, that's in the census. And now, you cannot go around asking every Canadian Do you smoke or not? So typically, in cases like that, what do you do is that you, you take a small segment of the population, maybe you, you asked 5000 people, maybe you asked 10,000 people, right? Do you smoke or not? And see what fraction of that small subset is smoke or not? Right? So that's the exact that's called a sample. And most of the time, right, so all remember in the last clip, I talked about opinion

surveys, electoral polls, right. So those are all examples of samples, you do not study the entire, entire population, right? What you look at is a small subset of them, and you ask and you collect data about them. So that's an example of a sample. Now, between population and sample rate, suppose you have data for the population or for the sample, what are the questions that arise? So one question so when you're the population data is how do you present the data? And how to analyze that data, analyze that data in the sense of which are the variables that you want to look at? What's the connection between different variables that you want to look at? How do you present those connections? How do you present whether the data for a particular variable? So these are the questions that surround data from population. Once you have, if you have data for sample, which is remember, a subset of the population, still have the same questions that how do you present an analyze the data? But you have this additional question now that how do you make inferences about the whole population based on this subset that you've studied? So you want to know something about this entire population? But you have only studied this subset here? So how do you make use of whatever you found in this subset to make I get some idea about what's happening in the entire population? And so that becomes an in very important question. When when you just have a sample and not the entire population data? And statistics? Is the science that tries to answer these sort of questions. So let me start by looking at this basic aspect that how to present data, and especially how to present data visually, right? How do you get a sense of data visually, because as they say, a picture's tells 1000 words. Right? So what are the important sort of pictures that you see in statistics? So let me start off with what's called qualitative or categorical data. And then I'll move on to quantitative data. So let's take an exit qualitative or categorical variable. Let's take a particular example. This is the vaccination status of COVID cases in Ontario on December 7 2021. So again, I went to the COVID-19 website for Ontario. And so where they what they give you is that of the new COVID cases in Ontario on December 7 2021, right, how many of them were were bought for unvaccinated people? How many were for fully vaccinated people? How many were for those with just one vaccine, so partially vaccinated? Are those with unknown status? Right. And this is the data so there were 424 of the cases were unvaccinated. 401 was fully vaccinated. 26 Partially vaccinated 77 Unknown status. So this is called a frequency distribution. So what you have here is this variable X with four categories, unvaccinated partially vaccinated, fully vaccinated, unknown status, right? So these are the four categories, right? And these are the frequencies there's the frequency of this category, and so on. So how would you present this data? So for qualitative and categorical variables, the one of the most commonest ways of presenting data is using what's called a bar graph or column chart. So let me show you what these are. So, so here, so this one here is a column chart. And this is a bar graph or a bar chart. So what it does, so in the column chart, you look at the X axis, it gives the four categories. And on the Y axis, it gives the frequency. Remember, for unvaccinated, this was for 24. So that's what it plots here. So partial vaccinated is 26. So that's the height of this column here fully vaccinated for a one and are known status 77. The same thing now if you lie them on the side, right, so now, on the Y axis are the four categories and on the X axis are the numbers. So then it is called a bar graph or a bar chart? And I'll show you later in an example where one type is better. Like maybe better than the other. But the basic idea is this that here, there is not much difference in this particular example, there is not much difference. But sometimes suppose you have a lot of categories, and the category names are pretty long, then it becomes very difficult to fit them on this X axis here. So, in that case you make use of this bar chart. So, what I've done is that here, I've created both of these in Excel. And in the next video, I'm going to show you how you can create these in Excel. Right? But before that, let's look at these, right? Because we have generated this pictures, what are these pictures telling us? There is some people you know, like pictures a lot more than numbers. So if we look at this pictures, so you can see that the unvaccinated that's the highest column, right? So that means this has the most number of cases. And then the fully vaccinated comes next, and it's very

close to the number of cases of non vaccinated, and this is those with partial vaccination. And this is those with unknown status. Right. So looking at this column chart, you see that the most number of cases is unvaccinated, but fully vaccinated comes a very, very close second. Now, I'm going to stop the clip here. But what I would like you to think about is what does this tell us so can you think about what inference can you draw from a picture like this? And it's the picture right about drawing that inference, or is it misleading in some way? So think about this, and I'll start with that. At this in the next clip.