

module_stat4_lecture10

📅 Thu, 2/17 4:18PM ⌚ 16:48

SUMMARY KEYWORDS

hypothesis, null hypothesis, hypothesis, alternate hypothesis, hypothesis testing, data, person, definitive statements, true, innocent, subsidized childcare, equal, question, basic idea, case, conjectures, test, bar, online, suppose

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

So now let's start on a slightly different segment of the module called hypothesis testing. Although the topic will be slightly different, but you'll see that the tools that we'll be using the same tools that we have been using so far in the module. So what's hypothesis testing about? So you'll see that oftentimes in the social sciences, people posit theories or hypothesis about how the world works or conjectures about what might be the impact of such and such policy or such and such change. So the question is, can we use data and statistical methods to actually check for those hypothesis check for those conjectures about the impact of some change some policy and so on. So this is the world of what's called hypothesis testing. So using data to statistically check if we have evidence for or against a particular hypothesis. So let me give you some examples, that suppose someone makes the statement that online learning has had an impact on student learning. So this is a hypothesis that the going from in person to online has had some impact on student learning. Or the sort of hypothesis is that policy A will have a positive or negative impact on outcome B. So what are actual examples of it? For example, people say that giving out free laptops can have a positive impact on children's education. Okay, what about subsidized childcare? I talked about this earlier, two rights were said some people say that subsidizing childcare will help both children and parents. It will probably increase employment opportunities for parents because they will they don't have the constraint of how also having to tend to young children at home. But that's a hypothesis. Question is, does subsidized childcare actually have a positive impact on children's development? And on parents outcomes? And economic policy? What about the raising the minimum wage? Does it have a positive or negative impact on the number of jobs? So these are all examples of hypothesis. Let me present a couple of other hypothesis. Someone can say that a country CO2 emissions are strongly related to its GDP. So this is a hypothesis about the connection between this variable and this variable. Another hypothesis that there's a strong connection between a city's crime rate and its unemployment rate. So again, this is a hypothesis between this variable and this period. So the question is, how do we test these sort of hypothesis or conjectures? So let's pick one of these right? And let me work towards the basic idea of, of, of testing of a hypothesis. Right? So let's take the hypothesis that online learning has had an impact on student learning. Okay, so that's my hypothesis. Just one thing, you'll notice that I am making very definitive statements. This is not a question that does online learning has had an impact on student

learning. So I've made a definitive statement that online learning does have an impact on student learning. And I'll connect this up to why we make such definitive statements. So suppose as a social scientist, you were thinking about what sort of data could be used to check this hypothesis. So let's say we have n students who have gone from in person learning to online learning, right? And let's say we have test scores of these n students before and after, like, well while they were doing in person versus while they were doing online, right, and let's look at did online learning have an impact on the test scores. So as a social scientist, this is the type of data methods that you use to try and answer hypothesis like this. Because obviously, you're not going to get completely perfect data to answer a question, right? What do you have to think about is what sort of data can you use to shed light on this question? So so what we're doing here is so we look at the score of a student online versus score of the same student, when do we teach learning? In person? Right? So let's look at the difference between these two scores, right? Let's call that x . And what we're interested in is what's the mean of x ? What's the expectation of x , right? And that we denote by the parameter μ . So this is this is our variable of interest, right? And what the hypothesis here says is the online learning has had an impact on student loan debt. So how do we cast it in the language of hypothesis testing? First, we set up what's called the null hypothesis, the null hypothesis, which is denoted by H_0 , right? So this says that μ is equal to 0. So what this what the typically the null hypothesis does is that it takes a skeptical view of the hypothesis that you're trying to that you're positing. Right, you take taking the skeptical view of that, that saying that online learning has had no impact on student learning. Right? So in that case, your null hypothesis is μ is equal to 0. And what you're checking this against is the alternate hypothesis, which says that online learning has had an impact on student learning. So the alternate hypothesis is that μ is not equal to 0. So, analogy to how to think about null versus alternate hypothesis, right? So suppose the police has arrested someone on the suspect of, of being involved in a crime of having committed a crime, let's think of a murder or some other gory crime, right? When this individual is taken to court, the basic premise in most legal systems or most Western and English based legal systems is that this person is innocent, unless proved otherwise, right. So in that case, the null hypothesis is innocent. Right, and the alternate hypothesis, so unless this null hypothesis is disproved, unless is disproved to be innocent, we cannot show that he's killed that. So the same thing here, we're going to take the skeptical view that this hypothesis is not true. Right? And then we're going to see the evidence, and then we're going to check whether this person based on the evidence whether we can still go with that premise that this person is innocent or not. Same thing here, based on the evidence from the data, can we still hold the null hypothesis? Or do we have to reject? So that will be the approach of, of hypothesis testing? Okay, so how do we go about this? Right? So how do we test this hypothesis that whether μ equal to zero or not, right? So what we're going to, again, appeal to is the central limit here. So remember, we have now the data on the difference in the test scores between online and the difference between online and in person? Right, so we have data on N students. And let's say we calculated the mean of that. Now, again, let's go back to the central limit theorem. Again, that's going to come to a rescue and that's why I've told you before that the CLT is a magical theorem for statistics, right? What the central limit theorem says is that this $\bar{x}$ right should be normally distributed around the mean. Right and with this variance now, if the H_0 hypothesis is if H_0 is true, right, so this mean should be zero. So that means this should be normally distributed around zero and σ^2/n . So that's what the central limit theorem tells you write. So what does it mean in terms of picture? In terms of the picture? What it means is that if you look at the distribution of $\bar{x}$, right, so this is going to be normally distributed like this, right? If this hypothesis H_0 is true, right, so then this $\bar{x}$ should be equal to zero. The second thing that we know is that for the normal distribution, right, so 95% of the observations, right? are plus minus two standard deviations away from the mean. Right? So 95% of the observations should lie minus two standard deviations below, or plus two

standard deviations above, right? They should lie in between this point at this point. So the basic idea of hypothesis testing is this, right? So almost 95% of the time, right? If this, if H_0 were true, right? 95% of the time, you should see your $\bar{x}$ lie in this range. But, but if your $\bar{x}$ lies here, or here, right outside the 95% range, it's unlikely that μ is equal to μ_0 . Again, if μ was equal to zero, you shouldn't see many observations here or here. Right? Because 95% of the observation should be here. Right? So if you see something here, or if you see something here, right, it's unlikely that your H_0 is true. So that's the basic idea of hypothesis testing. So how do we actually make a test out of this, right? So what it says is that if H_0 is true, right, there's 95% chance that $\bar{X}$ should lie between $\mu - 2$ standard deviation, and $\mu + 2$ standard deviation. So if H_0 is not true, this is zero, μ is zero, right? And you know, the standard deviation and this is what we have done before is to $s / \sqrt{n}$, right? And in this side, it should be plus two. So for square root of n , right? So 95% chance that $\bar{X}$ should lie between this point and this point. So what's our test? So test is this. See, you look at this particular interval, right? $\mu - 2s / \sqrt{n}$. So in this case, μ is equal to zero, in this case, μ was equal to zero, right? If $\bar{x}$ lies outside this right, so in this case, you know that it's very, very unlikely that H_0 is true. So in this case, you reject H_0 in favor of the alternate hypothesis. On the other hand, if $\bar{x}$ lies within this interval, right, then you cannot reject the null hypothesis. So this is the basic idea of hypothesis testing that again, let me go back to my normal diagram, right? If you find that the $\bar{x}$ lies within this interval, that's as you would expect, under H_0 . But if you find that $\bar{x}$ is outside that interval, if it's around here or here, then you're surprised, right? You're surprised in the sense that if the null hypothesis were true, if μ was equal to actually true, you shouldn't have been seeing these values. But now if you see a value like this, that means you're now hypothesis couldn't be true. So that's exactly what you're doing here is that if $\bar{x}$ is outside, that 95% range, right? You reject the null hypothesis in favor of the alternate hypothesis. If not, you cannot reject H_0 . So this again is the basic idea behind hypothesis testing. And in the next clip I'll do actual examples where with numbers to show how to operationalize this right. But a good analogy again with the, with the criminal case that I gave you before is that you started with the premise that this person is innocent. Right? Now, suppose this person cell phone was found at the scene of the crime. Plus a closed circuit, CCTV footage also showed him walking away with a blood stain shirt. Right now, if this person was innocent, but it's very, very unlikely that these two pieces of evidence would match up with his being innocent. Right? So in that case, there is a much higher evidence that this person is actually not innocent. The same thing here that you expect under the null hypothesis that the $\bar{x}$ should lie within that particular range. If it lies outside that means it's unlikely that that null hypothesis is true and therefore we will reject it. Okay, so that is the basic idea of hypothesis test.