

module_stat4_lecture4

📅 Thu, 2/17 4:18PM ⌚ 14:49

SUMMARY KEYWORDS

probability, central limit theorem, sample, bar, normal distribution, estimate, theorem, proportion, large numbers, chance, mew, tells, remember, called, draw, parameter, tails, clip, heads, population

SPEAKERS

Sumon Majumdar

Sumon Majumdar 00:05

So we are talking about using sample data to estimate the parameter of a population. And to estimate a parameter like μ , and if it's a quantitative variable, we use the mean of the sample $\bar{x}$. Or if it's a categorical variable, then we use the proportion who are of type one, or proportion who are of type zero. Which seems simple enough. Now the question that I raised, the last clip does that why is that a good estimate? Right? So this seems pretty simple to do. Why do we need even statistics for that? So what I'm going to do in this clip is give you a basis for arguing why this is a good estimator. And later on, taking it a step further, and telling us how much confidence can we have in such an estimate? So so the main question we are asking is, why is this $\bar{x}$ or P that we have from the sample data? Why is this good estimator? Now, the basis of this is probability. And we had done four modules on probability earlier. And hope you remember the material from that, because we're going to draw heavily on that material. So the first important theorem from probability in the context of statistics is what's called the law of large numbers. So what it says is that as n goes to infinity, so remember, n is the size of your sample, right? How many people are you sampling? So what this says is that as n goes to infinity, then $\bar{x}$, which is the mean of your sample, it approaches μ . This is the parameter that you're trying to find, right? This approach is mew with probability one. So this is called the law of large numbers. So in plain language, what it says is that as the size of the sample gets bigger and bigger, right, so as n goes to infinity, there is more and more and more chance that your estimate, which is the $\bar{x}$, this is close to the population parameter μ . So what this gives you is the basis that as you draw a bigger and bigger sample, there is a better and better chance that the number that you have that sample average $\bar{x}$, this is very close to what you're trying to find, which is the population parameter mean. So this is called the law of large numbers. So I'm not going to go through the go into proof of this. But I'm going to give you an intuition for why or based on something that we have seen before intuition for why this law of large numbers holds. So if you recall our, in our first module on probability, I've talked about an interpretation of probability. So suppose you toss a coin, this probability half that the it will be heads and half it will be tails. So one interpretation of probability was this that if you toss a coin many, many, many times, then the proportion of times that a heads comes up in a proportion of times that a tails comes up is half and half. So more generally, if you repeat an experiment, many, many times the fraction will go to the probability. And I've given you this diagram, if you recall, so this was something that I did in from a website called StatCrunch, where you can

simulate lots of events. So this is the case where I flipped a fair coin 1000 times and recorded the number of heads and the number of tails that comes up. So what this blue graph here tells you that so say, say, say I've done 200 draws right, what proportion of those 200 draws came up to be heads. If I keep repeating and go up to 400 How many of those 400 came out to be heads? How many of the 600 800 1000s? Right? So theoretically, we know that it's probably half, you get heads, and probably half you get tails. Now you see that initially, this proportion of heads, right, so this is bouncing it out. So, so this is the point five line, that's the green line, right? So initially, it's below this. But as the number of coin tosses becomes more and more and more, see, this line starts slowly approaching the green light. And beyond a certain point, it's, it's oscillating, it's going sometimes not going sometimes blue, but more or less, it's following this green line, which is point five. And that's the same idea of the law of large numbers that in general, as your sample size becomes bigger and bigger and bigger, this is like as the number of coin tosses becomes bigger and bigger and bigger, but what you get from the data, which is the proportion of times that this coin turns out to be heads, right, that comes closer and closer to what is the population parameter. So this gives you a basis that of the following statement that as your sample size becomes bigger and bigger, that means you're you get more and more confidence, or you get better and better chance that what you're estimating from your sample is coming closer and closer to the towards the reality in the population. Okay, so this is called the law of large numbers. Now, there is a second issue. The second issue is this, right? See, see you're interested in calculating new the mean value in the population, which you don't know. So you draw a sample. Now, so here's the population of Canada, right, you could have drawn a sample here are you could have drawn a sample from here, right? Or you could have drawn a sample from here, right? So depending on which sample you draw, right, the $\bar{x}$ that you get, right, so this will be different for this sample, this will be a different $\bar{x}$ for this sample, this will be a different $\bar{x}$ for this sample, this will be a different $\bar{x}$ for the sample, and so on. So for different samples, the $\bar{x}$ bar, that estimate that you do this will be different. So in fact, if I do the experiment, and I sample a few people, I'll get a different answer, compared to if you sample and you choose a different group of people. But at the end of the day, you just draw one sample and and you get an $\bar{x}$ bar. Right? So how much confidence do you have that $\bar{x}$ bar that you measured? This close to the to the what you're trying to measure, which is the Mew, how much confidence? Do you have that this $\bar{x}$ bar, right? As close to mew? Or is it very far away from me? Right? So what's the distance between this $\bar{x}$ bar which you have actually come up with and the mew that you're trying to find out? So the second theorem in statistics tries to answer this question. So this is called the central limit theorem. And this is a beautiful magical theorem. So what it says is the following. So remember, μ is the unknown parameter of the population that you're trying to find out. Right? And you have a sample of size n and from that you have estimated the mean $\bar{X}$ bar. So what the central limit theorem says is that as n goes to infinity, right, as your sample size becomes big, this $\bar{x}$ bar right, the thing that you estimated, will be normally distributed around you with the variance σ^2 over n so in shorthand, $\bar{x}$ bar the thing that you estimate it will be normally distributed with a mean of μ and the variance of σ^2 over n . So this will be the variance, and this will be the way. So this may not seem very much, right, why is so why is it so magical? Why did I use the term? Let me show this to you in a diagram. So, what this tells you what the central limit theorem says is that see, this is $\bar{x}$ bar, right, so, this is your estimate. Now, this $\bar{x}$ bar could be jumping around the place depending all over the place, depending on the sample that you draw, right. Remember what I said before, so, this $\bar{x}$ bar could be here, it could be here, it could be here, it could be here anywhere. But what the central limit theorem says is this the distribution of this $\bar{X}$ bar, right, so this will be centered around μ , right the thing that you're trying to find. And it will have this normal distribution. This bell shaped normal distribution. So what this is remember, again, this is from our probability module. And if you want to go back look at probability module four, where we talk about normal distribution a lot. So, this here, this

bell curve here is the PDF of the normal distribution. So, what this tells you is the relative chances of the different $\bar{x}$ soccer. And if you look at this, that you see where the highest relative chance of $\bar{x}$ occurring is very close to what you're trying to find out. And as you go further and further away from you, right, so, see this probability is decreasing. Right. So there is some chance that the $\bar{x}$ that you come up with makes turn out to be very far from you, right, but the relative probability of that occurring, which is given by this is going to be very small. Similarly, the $\bar{x}$ by luck, you could end up here, but the the chance of that happening is going to be very small that, where is the highest chance, the highest chance is in the regions very close to me. Because this is where you have the highest probabilities of $\bar{x}$ occurring. So, what the central limit theorem gives you is an idea of confidence that see, you draw a sample, you get an $\bar{x}$, right. So, this theoretically this $\bar{x}$ could lie anywhere, right. But what it says is that the highest chances are of it lying close to new, the one that you're trying to find out. And as you go further and further away from you, right, the chances decrease, which is exactly what you would like it, you'd like the chances to be highest close to what you're actually trying to find out. And the chances of it being very far away that falls. So that's why the central limit theorem is going to be super important for purposes of estimation. Now, so let me stop this clip here. And we will because the central limit theorem tells you that the $\bar{x}$ is going to be normally distributed, right in the next few clips, we are going to make use of the normal distribution a lot. So if you want to go back and refresh your memory about the normal distribution, this may be a good time to look at some of the clips there on the normal distribution, which you will find in probability modules four.