

module_prob4_lecture14

Fri, 2/18 12:51PM 12:31

SUMMARY KEYWORDS

module, 50th percentile, normal distribution, percentile, probability, continuous random variables, standard deviation, pam, score, norm, excel, concepts, cdf, 90th percentile, cover, continuous distributions, class, figure, connected, important

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:04

Hi, this is going to be the last video of this module. And sorry for going a bit longer in this module compared to previous ones. And that's primarily because the normal distribution is such an important, useful and widely used distribution, I wanted to cover all its essential properties for this module. So what we were looking at in the last module was looking at test scores, saying in SAT or a GRE exam, and if Pam and Tam took two different versions of the exam, how can we compare the test scores across them. And that was by using the technique of standardization, where we took the test score minus the mean, and divided by the standard deviation. So there's another corporate concept, especially in test scores, and also used in other places to compare between situations. So, one, so another way of looking at Pam's performance in her test is to see that okay, she scored 75 in the LSAT exam, in which the mean was 70, and standard deviation was 10. So what fraction of the class did worse than right? So where did Pam stand in the ranking of the class, so to speak? And this, we know how to do that. So we are going to calculate what's the probability that X tilde was less than or equal to Pam score. So this is the CDF at 75. And we can use Excel, say, NORM dot DIST. Remember that function that we used to figure out what this f must be? So this gives you what fraction of the class did worse than another way in which the SATs or GRE and other exam scores are often cast is in terms of percentiles. So what's a percentile? So if you look at the GRE scoring scale, so this is the score for the 25th percentile? 151 is the score for the 50th, percentile? 157, for the 75th 162 for the 90th, etc. So what does that mean? So what it means is that what is the score that you need to achieve? Right, so that 90% of the classes below you? So that will be the 90th percentile score. So what are you looking for is x such that f of x is equal to point nine. So so far, we've been looking at okay, let's take Pam score, what is the F connected with this? Now we're going to be looking at the reverse that suppose we want to find out a probability of point nine, what must be this corresponding value of x ? Right? So that will be the 98%. So what we're looking now is the reverse, we know the probability, we want to figure out what is this value affects. Similarly for undefined, the 75th percent that we're going to look at point seven, five, and we're going to look at what must be the x connected with that. And so an Excel allows you to do this as well. So this is the inverse, because now we know the probability we're trying to find x , the norm disk, we knew the x and we were trying to figure out the probability. So in Excel to do this, we use the norm in a function, right, we put in the probability that we want to

find and we put in the two parameters, the mean and the standard deviation. So let me show you an example of doing that in Excel. So here's my Excel file, the same one that we used before. Now I want to figure out the percentile of point nine, right, so the 90th percentile for Pam score, in which the mean is 70. And the variance is the standard deviation is 10. It's what I'm going to use is the NORM in NORM NORM in, that I'm going to put in the probability, which is going to be in I 3, right the point nine. And I want to put in the mean, which is going to be 70, and standard deviation of 10. So that gives me the the 90th percentile for Pam in the - test that Pam took is 82.8. Similarly, if you want to find the 50th percentile read, again, we're going to use the norm in in and it's going to be high for the mean is 70. And the third parameter is the standard deviation, which is stand that so the the 50th percentile is 70. So this is how you would figure out the various percentile 98/75, etc, percentile off a normal distribution. Let us now go back to our slide. So using Excel, we figured out that this 90th percentile here is going to be 82.8. The 50th percentile was set. If I for for for a normal distribution, the mean and the 50th percentile are the same. And which we also saw in the Excel sheet. So more general, any percentile P, right is the value of x corresponding to it. So if you were to figure out what's the percentile score connected with P, that is x. So and what that gives you is that what is the score, so that P fraction of the population is below that. So in Pam's case, the 90th percentile score was 82.8. So what that meant is that 90% of the class had scored below 82 points. And the common ones that we look at the 90th percentile, the 75th percentile, the 50th percentile. And in fact, the 50th percentile is very common, and it's called the median. Right? So median is the one that's correspond to the fifth point five, or the 50th percentile. So what this means is that half of the class is below the score. This in this case was 70. So half of the class is below this. So this is the idea of percentiles. So let me end the module by doing a quick recap of the things that we covered in this module. So began with a concept from the previous connected with the discrete random variables in the previous module that was of covariance and correlation. So this gave the idea of what's the connection or how strong is the relationship between two random variables. And remember, correlation, this runs from minus one to plus one. So when it's close to minus one, or plus one, that tells you that the relationship between the two random variables is strong. If it's close to minus one, that means the negatively correlated, and if it's close to plus one, that means they're they're positively correlated. Then we went on to the main segment of this module, which is continuous random variables. And I introduced the important concepts of the CDF and the PDF, the probability density function. And once we have that, then we derive the expectation and variance of continuous random variables. So for that, we had to appeal to integration, the concept of integration. And we did some examples of that. And then at the end of the module, I introduced some common continuous distributions, in particular, the uniform and the norm. And I spent quite a bit of time going over the normal distribution, because it's a super important distribution that you will see come up both in statistics, as well as in in lots of different areas of the social sciences. There are a lot more continuous distributions, like the exponential, Pareto, F, Beta, T. But we're not going to cover these in this module, because we have already kind of run over time. But you may see some of these distributions crop up in in different areas, or different subjects that you take. Now, the good thing is that all many of these distributions are already in built into Excel. And you can compute the values, the probability values, very similarly to how we computed that for the normal distribution, in Excel or in Google Sheets. So that's the end of our module on on probability, and continuous random variables. So I do hope you will practice the problems that were put up in the module, because that will give you a better understanding of the material that we've covered in this module. And most importantly, I do, I very much hope that you will remember the concepts that you learned here, especially things like the PDF, the CDF, and the normal distribution, because these concepts are very, very important for statistics for

things like regression and hypothesis testing, that we'll be covering in some of the modules and statistics, because probability, it's a little bit more puritanical than statistics, but it does form the basis of a lot of statistical concepts. So let's end the materials.