

module_prob4_lecture13

Fri, 2/18 12:51PM 14:13

SUMMARY KEYWORDS

variance, normal distribution, score, distributed, tilde, normal, squared, random variable, exam, subtract, distribution, divided, sum, variable, mew, test, infinity, standard deviation, variance sigma, multiplied

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:06

So we've been talking about the normal distribution. And as I mentioned, it's an extremely useful distribution, which is used in lots of things like height, weight, delivery times, test scores, IQ scores, in like in a wide variety of disciplines. So one thing that you may be worried about is that the normal distribution, the range, technically it goes from minus infinity to plus infinity. And we've been using things in cases like weight, height, etc. Of course, we doesn't go off to minus infinity to plus infinity. Right? So we looked at the example in the clicker question of a baby's weight, which the mean is 3.5 kg, right? So how added will be like three or two and a half to add most like one kg? So how do we account for that there's weights don't go off to minus infinity and so on. See, most of the cases, it doesn't matter. Because, as you know, 99.7% of the normal distribution observations, they lie within plus minus three sigma of the mean. Right. So in the base case, the the sigma is, is 1.5. So three sigma is 1.5. So that means the baby's 99.7% of the baby have weights between 3.5 minus 1.5, that's two cages or five cages. So almost all of the time, the normal distribution, that's why it works fine, because we don't have to worry about minus infinity and plus infinity cases. Now, there are also several properties of the normal distribution, which make it extremely useful for usage. And let me now talk about some of these. So one of these is that if x tilde is normally distributed with mean μ and variance σ^2 , right, now, if you add a constant to it, add a constant D to it. Or you multiply it by a constant like a . So what happens now to the mean is that the mean also gets transformed exactly the same way. Remember, the mean is view of the original x tilde. So now it becomes $a\mu + D$, and $a^2\sigma^2$. So we multiplied the where the random variable by a and added a D . So the mean also trunk gets transformed exactly the same way it becomes $a\mu + D$, but about the variance. So the variance is not impacted by D . And remember, this is a result that we also saw before when we talked about the definition of variance for a saving for a discrete random variable. So the only thing that changes now is that because the random variable gets multiplied by A , the variance also gets multiplied by A^2 . Right? So this new variable here, it's also normally distributed with mean $a\mu + D$ and variance equal to $a^2\sigma^2$. So what's an example of this right? So say x tilde is normally distributed with mean 3.5. And, and variance of point two, five, right? And we are interested in looking at the distribution of this random variable, which is two times x plus 10. So we know that this is also going to be normally distributed. What is it going to be? Its mean, say

it's going to be two times the regional mean, which is 3.5 plus 10. So that's seven plus 10, 17. What is its variance going to be? So see, this 10 doesn't impact the variance. The only thing that impacts is this two, it becomes now two squared, which is four times the original variance, which was point two five. So this is equal to one so now it's there. So this is a useful property of a normal distribution. Another useful property is that if you take two variables which are normally distributed, let's say x is normally distributed with mean μ_x and variance σ_x^2 . And y is another random variable, which has its own mean and its own variance. And let's say these two are independent, right, so one doesn't influence the other. So in this case, if you look at the sum, if you look at $x + y$, so this is also going to be normally distributed. And what is going to be the mean, it's just going to be the sum of the two old means the mean of X and mean of y . And the variance is also going to be very conveniently the sum of the two old variances, $\sigma_x^2 + \sigma_y^2$. So we really encounter something like this. So for example, right? Consider a husband and wife, right. And they both have their incomes are uncertain. One's income is normally distributed with husband's income as mean 2000, variance 100 wife's income is 2500 and variance 200. So what is the distribution of the joint so the joint income is the husband's income plus the wife's income, that's also going to be normally distributed. And the mean is just going to be the sum of the two means 2000 plus 2500 4500. So that's going to be 4500. What's going to be the variance, it's going to be 100, plus 200. So it's three. So this is an example where you use the sum of two normal areas. Similarly, you can talk about suppose the weight, like the family has three babies, right? And are has triplets maybe, right. And maybe you are interested in the sum of the weights of the three B's, each of them is normally distributed, the sum of the weights will also be normally distributed. And if those variables are independent, then the mean will just be the sum of the three means variance will be the sum of three variances. So that's a convenient property of a normal distribution. Now a lot of times, we look at a very particular normal distribution, which is quite useful. A normal distribution which has mean zero and variance of one. So mean zero and variance of one such type of a normal distribution is called a standard normal distribution. So the important thing about this is that it has mean zero and variance of one. So if you plot the standard normal, it'll have a bell curve, and with a mean of zero, right centered around the origin. Now, why is this useful and important? The reason is see if you have a normal variable, which is not standard normal, let's say it has some arbitrary mean μ , and some variance σ^2 , now if we take this random variable X and subtract the mean, so I'm subtracting a constant from it, now by the property that we just did a minute ago. So this, if you just look at $x - \mu$, this is also going to be normally distributed, right? So I've just subtracted a constant from it. What is that going to affect? It's going to take the mean, and you're going to subtract μ from it. So that's going to be zero, starting into impact the variance, so that's still going to receive a score. Now, on top of this, if I take this and divided by σ , right, so I'm dividing it by a constant, remember, again, the property this is also going to be normally distributed. So the new mean is going to be zero over σ . So that's zero. What's going to be the new variance? It's the new variance is going to be the old one. And one over σ^2 , right? It is a squared σ^2 . So if I am dividing this by σ , so this becomes one over σ^2 times σ^2 , C , this is going to be equal to one. So that means if you take the normal variable, subtract its mean, and divided by its standard deviation σ , right? It's going to have a normal 01 distribution. So this process is called standardization. So if you take a normal variable, subtract its mean and divided by its standard deviation, you get, you standardize it, right? And this is sometimes called the Z score. Where is this really useful? So this sort of thing comes up quite a lot when you're comparing across different situations. So for example, let me give you a quick example of this. Right? So sometimes people take this test, like a GRE score, a GRE test, or LSAT scores, right? Not all students take these tests at the same time. Maybe you take one version of the test, your friend takes another version and you cousin takes another version. And then the different versions.

Scores are random. And maybe they slightly differ. And they differ in the mean score, the spread of the scores, right? So how do you compare across those? So for example, take this example. So Pam scored 75, in an SSD examine which marks are normally distributed, with mean 70, and standard deviation of 10. And let's say firms friend or cause in town, she scored 85 in the SAP exam, but this sub examined a different mean a higher mean 75, the higher standard deviation 40. So question is who did that? Because these two are slightly different exams, because this is a higher mean, and a higher spread. This has a lower mean and a lower spread. So you will, if you want to compare between Pam and Tom, who did better in in the SAT, and we usually that's a problem that you encounter, always in like in tests, like the GRE, LSAT, and so on. It's what people do is they compute the Zed score. So what spams that score, she scored 75. So 75 minus the mean, which was 70x minus mu, divided by the standard deviation, which is 10. So that's five over a term, which is equal to half. What about time, law time got 85x minus mew, the mean in her exam was 75 divided by 40. So this is term over 40. So that's one four. So if you if you standardize the scores by accounting for the differences in the mean, and the variance, you see that Pam has done that, right. So Pam has done better relative to the other students who took her exam, as compared to tam although the term scored higher, but the mean in her exam was also higher, the variance was also higher. So relative to other students in in her exam. Pam score is better as compared to other students in her exam as as compared to so this is how you use standardization to compare across two situations.