4 PROBABILITY DISTRIBUTIONS

4.04 The variance of a random variable

After the mean, the second summary statistic you would like to know for a random variable is its variance, as a measure of spread. In this video I will explain how the variance of a random variable is calculated and also what happens to that variance when adjusting the random variable via addition or multiplication, or when combining different random variables.

The variance of a random variable X is defined as the expected value of the squared deviation of X from its mean μ.

If you want to calculate it on the basis of a probability distribution, it is the sum or integral over the squared difference between the values that the variable may take and its mean, times their probabilities. These are the equations for that calculation for a continuous and a discrete random variable.

Calculating the variance for a continuous random variable is difficult, because you would need to integrate this function, but for a given discrete variable it is less complex than it may seem.

Let's take an example. This discrete distribution gives the yearly risk that you would get involved in a traffic accident. The mean risks is 0.04, once in twenty five years. First you calculate the difference between the mean and each number of accidents, then square these differences, multiply them with the corresponding probabilities, and finally sum the result. The variance of the accident-risk appears to be approximately 0.06. If you would rather like to use the standard deviation to express the spread in the distribution, you can take the square root of the resulting variance.

Now let's see what happens with the variance of the random variable when that variable would be adjusted by adding a value A and multiplying with a value B. When you enter this transformation in the equation defining the variance, you see that it the constant A disappears, but that the factor B is being squared. Hence by adding or subtracting a value A to a random variable, it's variance does not change but when you multiply a random variable with a value B, its variance becomes the original variance times B-squared. The standard deviation, the square root of the variance, changes then with a factor B.

Time for an example. Did it ever occur to you that on a bight sunny day, people you encounter are more likely to greet you then on a gray rainy day? This is the distribution of the number of nods or smiles you can expect per minute when walking in a busy city on a gray day. A meagre average of 1.4 per minute, with a variance of 0.84. Now, at the same time of day and location but sunny weather – everyone seems to have become friendlier – two times as friendly to be specific . Here's the nod & greet distribution under sunny conditions - there's the same category of grumpy people who never nod, but for the rest you expect up to 6 smiles or nods per minute.

Theory tells that the average number of nods should become two times 1.4, so 2.8; and the variance should go from 0.84 to four times that value, which is 3.36.
Let's check this by calculating the variance for the new distribution. This table shows the steps. We take the difference between the mean and each number of smiles and nods, square the difference, multiply it with the probability, and then sum it – indeed 3.36.
By the way, do you know the unit that goes with this variance value? It's the unit of the random variable squared – so we're talking about smiles per minute – squared here.

Let's now consider what happens with the variance when two random variables are added or subtracted.

For random variables X and Y, the variance of their sum is the sum of their separate variances plus two times the covariance between X and Y. And the variance of the difference is an even more curious equation: it's the sum of their variances minus the covariance between X and Y.
These are more complete equations, which include multiplication factors A and B for X and Y respectively.

These equations apply to the situation where any two random variables are added or subtracted. And apparently it requires knowing the covariance between these two variables. Covariance information is often not available and therefore we will not consider the general situation here, but rather a more restricted case when the variables are not correlated. That's a lot simpler because the co-variance between uncorrelated random variables is zero, and these terms disappear from the equations. So in this case it doesn't matter anymore whether you add or subtract variables – the resulting variance is always the sum of the separate variances. And you can generalize the equation to any sum of random variables.

Another noteworthy aspect is that the standard deviation of the resulting sum of random variables is always smaller than the sum of the standard deviations for the separate random variables. It makes sense, some variability will cancel out when uncorrelated random variables are combined.

Let me summarize what I have explained in this video:

- The variance of a random variable is the sum or integral over the squared difference between the values that the variable may take and its mean, times their probabilities.
- Adding a constant to a random variable does not change its variance, but multiplication with a constant leads to a multiplication of the variance with the squared-constant.
- The variance of several uncorrelated random variables that are added or subtracted is the sum of their variances – mind you this only applies to uncorrelated random variables.
- The standard deviation is the square root of the variance so to get the standard deviation after manipulating a random variable, you apply the rule to the variance and then take the square root.