1 EXPLORING DATA

1.06 Variance and standard deviation

Tattoos are increasingly popular among football players. Imagine you want to know how much of their bodies football players cover with tattoos. The dot plots you see *here* represent the distributions of tattoo density (expressed by the percentage of the body covered with tattoos) in two football teams. You can see immediately that in the first team the tattoo density is much less variable than in the second team. This variability can be measured by, for instance, the range or the interquartile range. It can graphically be represented by a box plot. You see the relevant box plots *here*. In this video we'll discuss two other measures of variability that are used very often in statistical studies: the variance and the standard deviation. The huge advantage of the variance and standard deviation over many other measures of variability is that they take into account *all* the values of a variable.

Let's start with the **variance**. *This* is the formula of the variance. Let me show you how it works step by step. s^2 stands for the variance. *This* part of the formula means that from every observation (x) you have to subtract the mean value of that variable ($\bar{x}$). Next, you have to square all these values and add them up. The result is what we call the "sum of squares". In the following step you divide this "sum of squares" by the size of your sample (*n*) minus one.

Let's now apply the formula to our tattoo density example to see how it works in practice.
So, *these* are our "Team 2 data", and *this* is the formula. The first step is to compute the mean. I won't do that now because I assume that you already know how to do that. The mean of these values equals 15. The second step is to subtract the mean from every single observation. So, let's take the first value: 0. We subtract the mean from this value. That gives 0 minus 15 is -15. We do that for all our values in the sample. When we are finished doing that, we have completed *this* part of the formula.

Notice that you now have negative and positive numbers. This is not strange as the mean is the middle, or the balance point of these values. In fact, because the mean lies exactly in the middle, the negative deviations from the mean counterbalance the positive deviations from the mean, as a result of which the sum of the deviations equals 0. In other words: the sum of these values equals 0. For that reason we don't use the original deviations but the *squared* deviations.

That's the next step. We square all these computed values. So -15 squared is -15 times -15 is 225. We do that for all the other observations as well. According to the formula, we now have to add up all these values. After all, *this* is the "sum up" symbol. What we now have is the sum of the squared deviations, or, in other words, the **sum of squares**, which equals 639.74. We have to divide the sum of squares by *n* minus 1. The *n* in our case is eleven. So *n* minus 1 is 10. 639.74 divided by 10 equals 63.97. That's our variance!

The larger the variance, the larger the variability. That means: the larger the variance, the more the values are spread out around the mean. The first team, displayed *here*, has a variance of about 6.33. You can see that the larger variability of tattoo density in Team 2 that was already visible from the dot plots and the box plots is also represented by the larger variance.

An important disadvantage of the variance is that the metric of the variance is the metric of the variable under analysis *squared*. After all, we have squared the positive and negative deviations so that they don't cancel each other out. There is a very simple solution to get rid of this problem: we just take the square root of the variance. We call what we get the **standard deviation**. It can be seen

as the average distance of an observation from the mean. The larger the standard deviation, the larger the variability of the data. Because *this* is the formula of the variance, *this* is the formula of the standard deviation.

So, in our example, the standard deviation of Team 1 is the square root of 6.33. That equals 2.52. The standard deviation of Team 2 is the square root of 63.97. That equals 8.0. The standard deviation is the measure of dispersion that is used most often. However, in many statistical methods the variance plays an important role as well. In this video, you have learned that they are closely related, and that you can easily derive the one from the other.