1 EXPLORING DATA

1.05 Range, interquartile range and box plot

As you might have noticed, tattoos are increasingly popular among football players. The so-called "tattoo sleeve" in particular is rising on the football fields. A tattoo sleeve is what the name suggests: a sleeve of tattoos. You are interested in the question to what extent football players have covered their bodies with tattoos.

Imagine two football teams. What you see here are dot plots, representing the distributions of the variable "percentage of body covered with tattoos" in these two teams. The horizontal line represents this variable and the dots stand for the eleven individuals in each team. The players of Team 1 have covered about 10 to 20 percent of their bodies with tattoos. In the second team the players differ much more from each other in terms of their tattoo density. The percentage ranges from 0 to about 30 percent.

Thus, these teams strongly differ from each other. However, mode, median and mean are the same. In both distributions the mode equals 14.1 and median and mean equal 15. This indicates that in order to adequately describe a distribution we need more information than the measures of central tendency. In this video I will show you that we also need to have information about the variability or dispersion of the data. I will discuss two **measures of variability**: the range and the interquartile range. I will also discuss the so-called boxplot. A very useful graph that gives a good indication of how the values in a distribution are spread out.

The most simple measure of variability is the **range**. It is the difference between the highest and the lowest value. Let's look at our two teams again. The player in Team 1 with the largest tattoo density has covered 19.3 percent of his body with tattoos. The player with the smallest tattoo density has covered 10.8 percent of his body. The range 19.3 minus 10.8 equals 8.5. In Team 2 the player with the largest tattoo density has covered is body for 27.7 percent with tattoos, and the player with the smallest density for 0 percent. The range is therefore 27.7 minus 0 is 27.7. The range thus shows you at a glance that there is much more variability in Team 2 than in Team 1.

The range is a measure of variability that is easy to understand and simple to compute. However, in many cases it doesn't give a good impression of the variability of the data. The reason is that it only takes into account the extreme values. Look at *these* two distributions. They have the same range, but you can see immediately the variability in the second distribution is very different from the variability in the first graph.

Another measure of variability – the **interquartile range** – is a better measure of dispersion because it leaves out the extreme values. It basically divides your distribution in 4 equal parts. So, if your distribution looks like *this*, you divide the scores in such a way that the 25 percent of your lowest scores are below *this* score and the 25 percent of your highest scores are above *this* value. We also have 25 percent of our scores *here*, and 25 percent of our scores *here*. The values that now divide the distribution are called **quartiles**. *This* is the first quartile (**Q1**), *this* is the second quartile (**Q2**) and *this* is the third quartile (**Q3**).

As you can see, the second quartile divides the distribution in two equal parts. After all, 50 percent of the values is below this value, and 50 percent lies above this value. Q2 is therefore the same as the median. The interquartile range is the distance between the third and the first quartile, or, in other words, IQR equals Q3 minus Q1.

Let me show you how to compute it by going back to the tattoo density example. *This* is wat the distribution of Team 2 looked like. First, you look for the median, or, in other words, Q2. That's easy, it's the middle value. That's 15. You find Q1 by looking for the middle value of the values on the left side of the median. That's here: 8.7. You find Q3 by following the same strategy on the right side of the median. That's 19.3. Now, the interquartile range is Q3 minus Q1 equals 19.3 minus 8.7 equals 10.6.

The main advantage of the IQR is that it is not affected by **outliers** because it doesn't take into account observations below Q1 or above Q3. Yet, it might still be useful to look for possible outliers in your study. As a rule of thumb, observations can be qualified as outliers when they lie more than 1.5 IQR below the first quartile or 1.5 IQR above the third quartile.

There is one specific type of graph that is very useful when it comes to describing center and variability and detecting outliers. That graph is the so-called **box plot**. The box plot shows you at a glance Q1, Q2 and Q3, the minimum value that's not an outlier, the maximum value that's not an outlier, and the outliers.

This is a box plot based on the previous example. The box itself stands for the central 50 percent of the distribution. It goes, in other words, from the first quartile to the third quartile. The length of the box thus represents the IQR. The horizontal line inside the box is the median, or, in other words, Q2. *These* lines are called whiskers. They contain the other values except for the outliers which are displayed separately by means of dots. There are no dots here, so this box plot shows us that we don't have any outliers.

How do you decide how long the whiskers should be? Well, let's go back to the values in our example. We have detected Q2, Q1 and Q3, and the IQR. We know that values below 1.5 times the IQR below Q1 and above 1.5 times the IQR above Q3 are outliers. Our IQR is 10.6, so 1.5 times 10.6 equals 15.9. Q1 is 8.7. So all values lower than 8.7 minus 15.9 equals -7.2 are outliers. Such values don't exist, so we have no outliers on this side. Our minimum value is 0. That's the end of the whisker. Q3 is 19.3. So all values higher than 19.3 plus 15.1 equals 35.2 are outliers. We don't have values this high, so we don't have outliers on this side either. The end of the upper whisker therefore is equal to the maximum value, which is 27.7.

So, let's also take a look at the boxplot of Team 1. If we compare the two boxplots we see immediately that the variability within the two distributions differs strongly.

So remember, the center of a distribution only tells you one part of the story. For a more complete picture, also assess the variability of a distribution! A box plot shows important aspects of a distribution in a compact way, using the three quartiles, the outliers, and the range of the data after removing the outliers.