1 EXPLORING DATA

1.04 Mode, median and mean

Next to summarizing a distribution by means of graphs, it can also be useful to describe the center of your distribution. There are three main ways in which you can do that. By means of the mode, the median and by means of the mean. These three M's are often referred to as **measures of central tendency**.

Finding the **mode** is easy. It is the value that occurs most frequently. It is, in other words, the most common outcome. The mode is often used as a measure of central tendency if a variable is measured on a nominal or ordinal level. In *this* pie chart you can see which continent players in the main Spanish football competition come from. The pie chart makes immediately clear what the mode is: Europe. 70 percent of the players was born in Europe. Note that the mode here is Europe, which is the name of the category that occurs most often. The mode is *not* 70 percent. That's just the percentage of observations that fall in that specific category.

You can also have more than one mode. Imagine that there exists a football player that strongly divides football fans. Some people find him very sympathetic while others find him strongly unsympathetic. Let's name this player Franco Galtón. Imagine you have asked a representative sample of the Spanish population of 500 respondents what they think of Franco Galtón. Your respondents could indicate on a scale from 0 to 10 how sympathetic they think he is. 0 refers to very unsympathetic and 10 to very sympathetic. Let's say that *this* is the shape of the histogram resulting from your study. You can see that the Spanish population is strongly divided. Some find Galtón very unsympathetic and some find him very sympathetic. As you can see the distribution has two modes: 3 and 8. This is clearly a bimodal distribution.

The second measure of central tendency is the **median**. The median is nothing more than the middle value of your observations when they are order from the smallest to the largest. Imagine you have also asked 7 of your respondents what they think of another famous football player, named Tomás Bayez. Let's assume that *this* is the data matrix resulting from your study. The mode here is 8, the value that occurs most often. To compute the median, we first have to order all values from low to high. *This* is the result. Then we have to pick the middle value. So the median is 8.

It is slightly more complicated if we have an even number of cases instead of an odd number of cases. Imagine we haven't asked 7, but 8 people what they think of Tomás Bayez. *This* is the data matrix. And this is the order of the values from low to high. However, in this case there is no single middle value… How do we solve that problem? Well, we just take the average of the two middle values. That's 7 and 8 divided by 2 equals 7.5. The median in this case is 7.5. Notice that the median divides the distribution into two equal parts: fifty percent of the values lies below the median, and fifty percent above the median.

The third measure of central tendency is the most often used one, and also the one you most probably already know quite well: the **mean**. The mean is the sum of all the values divided by the number of observations. *This* is the formula with which you can compute the mean. It looks more complicated than it is. The formula tells you that the mean of variable X (symbolized as $\bar{x}$) equals the sum of all the values of X divided by the sample size (which is symbolized by *n*).

To give an example, let's again use the study on Tomás Bayez. This was the data matrix. The formula tells us to first sum all the values. That's 6 plus 7 plus 7 plus 8 plus 8 plus 8 plus 9. That equals 53. We

now have resolved *this* part of the formula. We now have to divide by *n*. The sample size in this study is 7. So 53 divided by 7 equals 7.6. The mean is 7.6.

You can think of the mean as the balance point of your data. Imagine we would place weights on a balance, one for each observation. Then the mean is the point on the balance where the total weight on the one side exactly equals the weight on the other side.

You're quite familiar now with the three M's and you can easily compute the middle of a group of scores in various ways. But when should you report which measure of central tendency? That partially depends on the measurement level of your variable. If it is nominal, it is impossible to compute the median or the mean. Think about it: you cannot apply numerical operations on nominal variables, nor can you order them. The only appropriate measure of central tendency when a variable is nominal is the mode.

But what to do in case of a quantitative variable? Imagine you're sitting in the canteen of a football club in your home town and you would like to compute the mean and median income of all persons present. That's you, five other guests, and the bartender. *This* is the data matrix. The mean is around 35,000. The median is exactly 35,000. They're pretty close to each other and it doesn't matter much which one you use to describe the center of your distribution. But now imagine the famous football player Franco Galtón walks into the canteen. Say he gets about 70 million per year. The median increases slightly to 36,000. The mean, however, becomes more than 8 million now…

We say that Franco Galtón is an **outlier** in this distribution. He earns much more than all the other people present, and his income exerts a disproportional effect on the mean income. In this case it might be argued that it makes more sense to compute the median than the mean to describe the center of the distribution.

Let me briefly summarize what you've learned in this video. To describe the center of a distribution you can use three measures of central tendency: the mode, the median and the mean. If your variable is categorical you use the mode and if it is quantitative you employ the median or the mean. Go for the median if you have influential outliers or if the distribution is highly skewed. If that's not the case, go for the mean!