1 EXPLORING DATA

1.03 Graphs and shapes of distributions

Researchers often want to summarize the data they have. They can do that, for instance, by means of a frequency table. In this video I will show you how you can use frequency tables to build informative graphs. I will also discuss the possible shapes that the data distributions in these graphs could take.

Imagine I studied where the football players in the main football competition in Spain come from. This frequency table could be the result. You can see that 280 players come from Europe, 16 from North America, 56 from South America, 32 from Africa and 16 players from Asia. I have also added the relevant percentages. You might want to present these results by means of a graph. Let me show you two possible ways in which you could do that.

What you see here is a **pie chart**. The categories of the variable you would like to summarize are displayed by means of slices of a pie. In a pie chart the surface of the slices represent the percentages of observations in each category. You can see at a glance that almost three quarters of all the football players come from Europe. Another way to summarize the same data is with a **bar graph**, which also shows you very clearly how the data are distributed over the various categories of your variable. The height of the bars represent the percentages of observations in each category.

Both types of graph have advantages and disadvantages. An advantage of the pie chart is that you can see immediately that about 75 percent of the players come from Europe. You cannot discern that information *that* easily from the bar chart without making a few calculations. On the other hand, however, the exact number of players in each category is not easily retrieved from a pie chart; while in a bar chart you can see easily, for example, that a little over 50 players come from South America.

A bar graph has advantages over a pie chart if the number of categories of a variable increases. Imagine, for instance, that you don't want to know which *continent* the football players come from, but that you want to know in which particular *country* they were born. *This* pie chart displays the result. As you can see it is very messy. You might like all the colors and small pies for aesthetic reasons, but all this information does not make the graph easy to understand. In this case it would make more sense to make a bar graph. *This* graph contains a lot of information as well. Yet, I would argue that it is much better digestible than the too colorful pie chart you saw before.

Up till now we have talked about a categorical, or more precisely, a nominal variable. How can we summarize data if we are dealing with a quantitative variable? One possibility is with a **dot plot**. The idea is easy. Imagine you have information about the physical height of 10 football players, expressed in centimeters. *This* is the data matrix. First you draw a horizontal line and label the possible values on it, in regular intervals. Like *this*… Next, for each observation, you place a dot above its value on the horizontal line. So *here*, *here*, *here* and *here*. That's it!

You can imagine that a dot plot is useful when you have only a couple of observations. However, it becomes messy when you have a large sample. With a sample of one hundred players, it would look something like *this*. When we have many observations, researchers therefore usually make use of another type of graph: the **histogram**. *Here* you see one. A histogram is similar to a bar graph in the sense that it uses bars to portray the frequencies or relative frequencies of the possible values of a variable. However, there is one important difference. That difference is that the bars in a histogram

touch each other. This touching represents that the values of an interval/ratio variable represent an underlying continuous scale.

Say we are interested in the body weight of Spanish football players. If we have very detailed measures of weight like 83.9 or 74.5 kilograms, it doesn't make sense to draw a separate bar for every single value. Instead, we construct intervals. In this graph we have 10 intervals of 5 kilograms. The first interval ranges from 47.5 kilograms to 52.5 kilograms. 50 is displayed because it is the middle of that interval. There are no fixed rules for how many intervals to make. However, it is important that the intervals have equal widths. So, in this case, always 5 kilograms. You can see at a glance that most players weigh around 75 kilograms. You can also see that a weight of less than 60 or more than 90 is rather exceptional.

As you can see, this histogram has a particular shape. It **has the shape of a bell, has one peak, and is approximately symmetric**. You will encounter such distributions very often. But not all histograms have this shape. A histogram can also be **skewed** to the left or to the right. A skewed histogram is not symmetric because one side of the distribution stretches out further than the other. *This* histogram is skewed to the left, and *this* one is skewed to the right. A variable that might have a right-skewed histogram is the annual income of the football players in the Spanish competition. There won't be many players with a very low income compared to the average income of the players. However, there will be some players who earn much more money than the majority of the players. For that reason there is a longer right tail.

A histogram could also have two peaks instead of one. Imagine a football match between two teams of six-to-eight-year-old players. After the match all children and their parents go for a drink in the canteen. You are interested in the question how old the people in the canteen are. Well, the histogram of the variable age might in this case well have two peaks. After all, those present in the canteen are children between 6 and 8 years old, *and* their parents which are, most likely, somewhere between 30 and 40 years old. You will therefore probably have a peak around 7 and a peak around 35 years old. We say that this variable has a **bimodal** instead of a **unimodal** distribution.

Okay, the most important lesson to take home from this video is that it is always a good idea to summarize your data by means of graphs. If you're dealing with nominal or ordinal variables you should make a pie chart or a bar graph, and if your variable of interest is an interval/ratio variable, you should make a histogram. And don't forget to look at the shape of your histogram. Is it bell-shaped and symmetric? Is it unimodal or bimodal? Is the distribution skewed? Assessing the shape of a distribution is of essential importance because it could affect the statistical methods you are going to employ later on.