1 EXPLORING DATA

1.01 Cases, variables and levels of measurement

Imagine you're very interested in football – you know, that sport that some of us like to call 'soccer'... You are that person who wants to know *all* the details, like: how many goals were scored by some player, how many games were won by a particular team, or how many penalties were stopped in a certain competition. In this video I will explain how improving your knowledge of statistics, could make you a real expert on football – or any other kind of sport.

The number of scores goals, won games and stopped penalties are all pieces of information that can be thought of in terms of **variables** and **cases**. Variables are characteristics of something or someone and cases are that something or someone. Let me be a little more specific. Imagine you are interested in some characteristics of football players belonging to your favorite team. Of every single player you want to know his or her body weight, hair color, age, and the total number of goals scored during the most recent competition. All these player characteristics are variables. The players themselves are cases.

Another example. It could be the case that you are not so much interested in the characteristics of individual *players*, but in the features of the *teams* these individuals play for. For instance, you might want to know about every Spanish team in which city it is based, what the main colors of the shirts are, and how many goals the team scored in the last year. These characteristics are variables again. However, the cases here are not individual football players, but the teams these individuals play for.

In a study, cases can thus be many different things. They can be individual football players and football teams. But they can also be, for instance, companies, schools or countries. Every characteristic of a case can be called a variable, as long as it meets one essential criterion: it needs to vary. What does that mean? Let's go back to the example with teams as cases and look at the variable "city where the team is based". You focus on *every* Spanish team, so there will be many different cities. One team comes from Barcelona, and other teams come from, for instance, Madrid, Valencia or Sevilla. We have, in other words, variation.

Let's now focus on another characteristic: not the city but the *country* where the team is based. For every single team this will be Spain – the teams are, after all, Spanish teams. This means that there is no variation here. Not a single team will be from another country than Spain. For this reason we call this characteristic not a variable, but a **constant**.

You can probably imagine that we can have many, many different kinds of variables, representing strongly diverging characteristics. For this reason, and also for other reasons that I will discuss later, it is of essential importance to distinguish different **levels of measurement**.

The most simple level of measurement is the **nominal** level. A nominal variable is made up of various categories that differ from each other. There is no order, however. This means that it is not possible to argue that one category is better or worse, or more or less than another. An example is the nationality of the football players. The various categories – for instance, "Spanish", "French" or "Mexican" – differ from each other, but there is no ranking order. Another example is the gender of the football players, or the city the football teams come from.

The second level of measurement is the **ordinal** level. There is not only a difference between the categories of a variable; there is also an order. An example is the order in a football competition. You know who is the winner, you know who came second and third etcetera. However, by looking at the

order you don't know anything about the differences between the categories. You don't know, for example, how much the number one was better than the number two.

Both nominal and ordinal variables can be called **categorical** variables. The next level of measurement is the **interval** level. With interval variables, we have different categories and an order, but also similar intervals between the categories. An example is the age of a football player. We can say that a player of eighteen years old differs from a player of sixteen year old in terms of his or her age. In addition, we can say that this player is older. But we can *also* say that, in terms of age, the difference between an eighteen-year-old player and a sixteen-year-old player is similar to the difference between a fourteen-year-old player and a twelve-year-old player.

The final level of measurement is the **ratio** level. It is similar to the interval level, but has, in addition, a meaningful zero point. An example is a player's body height measured in centimeters. There are differences between the categories, there is an order, there are similar intervals, *and* we have a meaningful zero point. A height of zero centimeters means that there is no height at all. (Note that we cannot say that age has a meaningful zero point because an age of zero does not mean that there is no age… Age therefore is an interval variable.)

Interval and ratio variables are what we call **quantitative** variables because the categories are represented by numerical values. Quantitative variables can also be distinguished in **discrete** and **continuous** variables. A variable is discrete if its possible categories form a set of separate numbers. For instance, the number of goals scored by a football player. A player can score, for instance, 1 goal or 2 goals, but not 1.21 goals.

A variable is **continuous** if the possible values of the variable form an interval. An example is, again, the height of a player. Someone can be 170 centimeters tall and 171 centimeters tall. But also, for instance, 170.2461. We don't have a set of separate numbers, but an infinite region of values.

Why is it so important to distinguish these various levels of measurement? Well, because the statistical methods we employ to analyze data depend on the level on which our variables are measured.

However, in practice the distinctions sometimes get blurred. For instance, for many statistical analyses the difference between the interval and ratio level is not that important. Moreover, many statisticians argue that if you have an ordinal variable measured on a scale with ten categories or more, you are allowed to analyze this variable as if it were quantitative. An example is a survey question that asks: "On a scale from zero to ten, how good would you say player X is?" Formally, this is an ordinal variable, but in practice you are allowed to cheat and to treat it as if it were a quantitative one.

To conclude. How does all this information make you a better expert on football?
Well, thinking about players, teams and competitions in terms of cases, variables and the levels of measurement of these variables, makes your knowledge about football more structured.
To become even more of an expert, do not hesitate and watch the next videos too!