1 EXPLORING DATA

1.02 Data matrix and frequency table

If you're conducting a study, it makes sense to think about your data in terms of cases and variables. Cases are the persons, animals or things you're studying, and variables are the characteristics of interest. In this video, I will discuss how you can order and present your cases and variables.

Imagine you are interested in the "Primera División", the top football competition in Spain. The cases you're interested in are individual football players within the league, and the variables you focus on are age, body weight, goals scored, team membership and hair color. The best way to order all this information is by means of a **data matrix**.

This is such a data matrix. The data matrix is the core element of every statistical study. It is nothing more than an overview of all your cases and variables. The cases are displayed in the rows. They range from "player one" to "player 400". You can see that no names are displayed, which means that the names are anonymized here. The variables are displayed in the columns. We have, as you can see, 5 variables: age, weight, goals scores, team membership and hair color. The values that are displayed in the cells of this table are usually called observations. 80.3 here means that Player 7 weighs 80.3 kilograms. The value 8 here means that Player 3 has scored 8 goals.

What you see here is not the complete data matrix. It is only a part of it. The complete matrix does not fit on this screen, because it contains 400 rows. We have, after all, 400 players. By means of *these* dots here I have made clear that I have left out a part of the matrix.

Let's see if our data matrix does not contain strange values. …. *Hey*… when we look at player 24, we see no value for weight. And in the next row, age is missing. So we don't know the value for every case-variable combination. For now we have just included these incomplete cases, but we might have to remove them if a subsequent analysis requires a complete data matrix.

You need the data matrix for all your statistical analyses. However, you usually do not present your complete data matrix to other people. The reason is that a data matrix is often huge – in our case we have 400 rows – and doesn't give a clear overview of the statistical information contained within the data matrix. When we present the information in our data matrix to others we therefore often make use of summaries of data in the form of tables and graphs.

Imagine you want to summarize the information you've got about the hair color of the players in the Spanish football competition. A good way to do that is to make a **frequency table**. A frequency table shows you how the values of a variable are **distributed** over the cases. A frequency table is nothing more than a list of all possible values of a variable, together with the number of observations for each value.

Here's an example based on the variable hair color. We can distinguish 4 categories: blond, brown, black, and other. *This* is the frequency table. You can see that 76 football players have blond hair and 160 players have black hair. Note that these values add up to 400, we don't have any missing data for hair color. We can also express the relative frequencies by means of **percentages**. In the second column you see the percentages. You can see at a glance that 7.5 percent of all players has another hair color than blond, brown or black. 19 percent of the players has blond hair. You get the value 19 here by dividing 76 by 400 and multiplying that with 100.

Sometimes researchers use **cumulative percentages**. It is easy to compute them. Cumulative percentages are nothing more than the percentages in every category added up. So you can see here that 19 plus 33.5 percent equals 52.5 percent of all players has blond or brown hair.

In this example we talked about the *categorical* variable hair color. What if we are dealing with a *quantitative* variable? Take weight, for instance. It doesn't make sense to compute percentages for every specific value of weight because then we would end up with a countless number of categories. The frequency table would show, for instance, that two persons have a weight of 65.3 kilograms, one person a weight of 65.4 kilograms, etc. That doesn't give you a good overview because it barely tells you more than the original data matrix. What researchers usually do to solve that problem is building new *ordinal* categories, by using intervals.

You could say for instance that the first category contains those players who weigh less than 60 kilograms, the second those who weigh between 60 and 69.9 kilograms, the next one between 70 and 79.9, the following one between 80 and 89.9 and the final one 90 and more kilograms. This way you lose information. But an advantage is that you get a much better overview.

We say that you have **re-coded** the variable. The variable weight was a quantitative variable, that you've turned into an ordinal variable with only five categories. It is easy to recode quantitative variables into ordinal ones. However, the other way around is impossible: you cannot recode ordinal variables into quantitative ones.

So, what do you know now? You use a data matrix as the source of all your statistical analyses. It is the overview of your data. However, if you want to present your findings to other people, you make use of summaries of your data. One very good way to summarize is by making frequency tables. If necessary, you can recode your quantitative variables into ordinal ones.