2 CORRELATION AND REGRESSION

2.01 Crosstabs and scatterplots

Many people like eating chocolate. Yet most people are somewhat cautious with their chocolate consumption, because it might well be the case that eating a lot of chocolate increases your weight. In this video, I'll talk about how we can display a relationship between two variables using tables and graphs. This can be very useful to help you discover if two variables are correlated or not.

Let us investigate the relationship between eating chocolate and body weight further. Suppose I have selected 200 female students at my university who are all 1m70 tall. This way height is a constant and cannot account for differences in body weight (or chocolate consumption). I asked the students to report their body weight and their weekly chocolate consumption. They could choose between the categories 'less than 50 kilograms', '50 to 69 kilograms', '70 to 89 kilograms' and '90 kilograms or more'. They could indicate their chocolate consumption by choosing 'less than 50 grams per week', 'between 50 and 150 grams per week', and 'more than 150 grams per week'.

Here are the results. What you see here is a **contingency table**. A contingency table enables you to display the relationship between two ordinal or nominal variables. It is similar to a frequency table. But the major difference is that a frequency table always concerns only *one* variable, whereas a contingency table concerns *two* variables. In our study we have two variables: body weight and chocolate consumption. The table shows that we have 33 individuals with a body weight of less than 50 kilograms. 27 of them eat less than 50 grams of chocolate per week. You can also see that 90 individuals eat between 50 and 150 grams of chocolate per week. 7 of them weigh 90 kilograms or more.

In this form the table does not tell you much yet about the correlation between the two variables, because the columns and rows contain different numbers of cases. It provides more insight when you compute percentages. In *this* case we compute column percentages. This means that for every cell we compute the percentage of cases in that cell compared to the total number of cases in the corresponding column. So, for instance, in *this* cell we have 24 cases. The total number of cases in *this* column is 60. So the percentage is 24 divided by 60 times 100. That equals 40 percent. When we do that for every cell, *this* is the result.

We can also express these percentages as proportions. 45 percent then becomes 0.45. 38 percent becomes 0.38. We call these proportions **conditional proportions**, because their formation is conditional on another variable. In this case chocolate consumption. We can also ignore the information we have about chocolate consumption and use the counts in the margin of the table. These are **marginal proportions**. For instance 33 divided by 200. That makes 0.17. This proportion tells you that a proportion of 0.17 (that's the same as 17 percent) of the respondents in this study weighs less than 50 kilograms.

So what does this mean? Of those who eat more than 150 grams of chocolate per week, 56 percent weighs 90 kilograms or more. Of those who eat less than 50 grams of chocolate, only 5 percent weighs 90 kilograms or more. Also, of those who eat less than 50 grams of chocolate, 45 percent weighs less than 50 kilograms, while of those who eat more than 150 grams of chocolate only 2 percent weighs less than 50 kilograms. These percentages show that those who eat more chocolate are also more likely to weigh more and those who eat less chocolate are also more likely to weigh less. In other words: the percentages show that there is a correlation between chocolate consumption and body weight.

A contingency table is useful for nominal and ordinal variables, but not for quantitative variables. For quantitative variables a **scatterplot** is more appropriate. Suppose that instead of providing categories, I asked the 200 women to give me their exact body weight; for instance 65 or 72 kilograms. Suppose I also asked them to tell me how much chocolate they eat every week. That could be, for instance, 64, or 99 grams per week. Now I have much more *precise* information than before. The best way to display the relationship between the *quantitative* variables chocolate consumption and weight is with a scatterplot.

To make a scatterplot we draw two lines, which we call axes. We call the horizontal axis the x-axis; here we display the independent variable. The vertical axis is called the y-axis, which we use to represent the dependent variable. If there is no clear distinction between dependent and independent, the placement on the y-axis and x-axis is a matter of choice. In our case, the independent variable is chocolate consumption and the dependent variable is body weight. Imagine that our study shows that the lowest amount of chocolate eaten is equal to 0 grams per week and the highest amount is 700 grams per week. We display these values on the X-axis. Similarly, the minimum value when it comes to body weight is 40 kilograms and the maximum value is 110 kilograms. Next, we display every individual in this figure. For instance, *here* is one person that eats 49 grams of chocolate per week and weighs 65 kilograms. Another individual eats 134 grams of chocolate and weighs 73 kilograms. That’s *here*. We do that for all individuals in our sample. And *voilà*, there’s the scatterplot! The scatterplot shows you at a glance that there is a relationship between chocolate consumption and body weight. The more chocolate you eat the higher your body weight.

So what have you learned? Not that chocolate consumption and body weight are correlated. I think most of you were already aware of that. What you have learned in this video is that we can display relationships between two variables by means of tables and graphs. When the variables in a study are measured on a nominal or ordinal level we use a contingency table and when they are measured on a quantitative level we use a scatterplot.