

Table 8.4 Cross-tabulation of the variables *Level of socialization with peers* and *Intention to quit this job*

	Intention to continue with the present job	Intention to find another job soon	Totals
High level of socialization with peers	195	45	240
Low level of socialization with peers	40	20	60
Totals	235	65	300

A table such as Table 8.4 is called a **two-way table**, or a **contingency table**, or a **cross-tabulation** of the two variables. We can read in it that we have the answers for 300 employees, of which 240 have a high level of socialization with their peers, and 60 a low level of socialization. Of these same 300 people, 235 do not plan to leave their jobs for the time being, and 65 wish to find another job soon. The number written in the lower right corner is the *grand total*; the other totals are called *marginal totals*.

Can we determine, on the basis of that table, that there is some kind of link between the fact that people do not socialize with their peers and their desire to leave this job? In order to answer this question, it may be helpful to compute some percentages. We will compute the row percentages, that is, the percentages within the categories of socialization with peers. The results are shown in Table 8.5.

Table 8.5 Cross-tabulation of the variables *Level of socialization with peers* and *Intention to quit this job*

	Intention to continue with the present job	Intention to find another job soon	Totals
High level of socialization with peers	195	45	240
Percentage within Level of socialization with peers	81.25%	18.75%	100%
Low level of socialization with peers	40	20	60
Percentage within Level of socialization with peers	66.6%	33.3%	100%
Totals	235	65	300

We can now notice the following:

- Among those who have a high level of socialization with their peers, 18.75% plan to find another job. This is a little less than 1 person out of 5.
- Among those who *do not* have a high level of socialization with their peers, 33.3% plan to find another job. This is 1 person out of 3.

**Remark.** The question asked in the survey could be:

Q. Would you say you have a high or low level of socialization with your peers at work? (check one)

A. High level: \_\_\_\_\_  
Low level: \_\_\_\_\_

But this is not a very good question, because there is no uniform definition of what is a high level or a low level. Instead, there could be a series of indicators, represented by questions such as:

Do you take your lunch with your peers or alone?  
Do you walk or drive home with some of them?  
Do you phone some of them during the weekends?  
If you have a problem with the boss, would you trust them enough to seek their advice?  
Etc.

On the basis of the answers to these questions, the researcher would divide the respondents into two groups: those who display a high level of socialization and those who don't. The criterion for classification could be something like: those who answered Yes to most of these questions will be classified as having a high level of interaction.

Here, the *concept* that we are trying to observe is the *level of socialization with peers*, and all the other variables (having lunch with them, calling them, etc.) are *indicators* of that concept. (Review Chapter 1 on these notions.)

We therefore notice a big difference between those who socialize with their peers and those who do not. In the latter category, a larger percentage of individuals plan to leave their job. We can say, therefore, that:

**Individuals in this sample who do not socialize with their peers are more likely to want to find another job than those who do socialize with their peers.**

The preceding sentence illustrates the fundamental aspect of statistical association between two categorical variables: People who are in one of the categories of the first variable are *more likely* to find themselves in a given category of the second variable. Thus we can conclude:

**There is a statistical association between the variables *Level of socialization with peers* and *Intention to quit this job*.**

Keep in mind, though, that it does not follow from that conclusion that the level of socialization is the *cause* of the intention to quit. It could well be the other way around. Or both variables could result from a third reason not presented in this table, such as: this place of work is in a remote area, far from people's houses. We will come back to the interpretation of the statistical association later in this chapter.

There is another way of looking at the statistical association described above. Instead of looking at the percentages within the levels of socialization, we could look at the percentage within the categories of the variable Intention to quit this job. We would get Table 8.6.

Table 8.6 Cross-tabulation of the variables *Level of socialization with peers* and *Intention to quit this job*

	Intention to continue with the present job	Intention to find another job soon	Totals
High level of socialization with peers	195	45	240
Percentage within Intention to quit job	83.0%	69.2%	
Low level of socialization with peers	40	20	60
Percentage within Intention to quit job	17.0%	30.8%	
Totals	235	65	300
	100.0%	100.0%	

We can now make an analysis similar to the one we made above. Among the people who plan to continue working at the same place, 83% maintain a high level of socialization with their peers. But that percentage drops down to 69.2% among those who wish to find a job somewhere else. Thus, we can say that *the individuals of this sample who do plan to stay in this job tend to socialize with their peers at a higher level than those who plan to leave*. Again, this indicates (or confirms) that there is a statistical association between the two variables.

Note that the percentages written in the two tables above are called either:

- row percentages** if they add up to 100% horizontally, across the cells of one row, or
- column percentages** if they add up to 100% vertically, across the cells of one column.

You will learn in Lab 10 how to produce similar tables with SPSS. Keep in mind that we are only talking about statistical associations, not about causes. It does not follow from the existence of a statistical association that one of the variables is the cause of the other.

## The Case of One Quantitative and One Qualitative Variable

Suppose now that we want to analyze the statistical relationship between one quantitative and one qualitative variable, for instance Income (quantitative) and Sex (qualitative). Several options are offered to us. The simplest is to compute the average of the quantitative variable separately for each category of the qualitative variable.

### Example

The average income for a sample of 1500 people, consisting of 800 men and 600 women, is \$19,400 a year. Suppose that the average income for women and for men separately is given by:

Average income of men: \$23,400

Average income of women in that sample: \$17,300

This would mean that there is a large difference between the incomes of men and women. The income of men is  $(23,400 - 17,300) / 17,300 \times 100 = 35.2\%$  higher than that of women.

This means that **there is a statistical association between the variables income and sex** for the individuals of that sample (we are not generalizing to the whole population yet). However, the preceding statement does not mean that sex is the *cause* of the difference in income. All we can say for the time being is that women make less money than men do. The *interpretation* of that difference is another matter. It could be due to discrimination (direct or systemic), it could be due to some other intervening variable (if, for instance, the women of this sample tended to be younger than the men, and therefore have less working experience) or some other cause.

Finding the average for men and for women separately is not the only way to establish the existence of a statistical association. Another method would be to recode *income* into three categories: high, intermediate, low, and then treat both variables as categorical variables. In SPSS Lab 5, you have seen in detail how to illustrate the difference between the incomes of various groups graphically with box plots. SPSS Lab 11 shows how to compute statistical measures for each group separately.

### Ordinal Variables

There are specific methods for establishing statistical association between ordinal variables. Such methods take into account the ranking of each individual on one of

the variables in comparison to his or her ranking on the other variable. They will not be treated here. Ordinal variables are often treated as quantitative variables and correlations are computed. The results of such computations are sometimes difficult to interpret.

### Statistical Association as a Qualitative Relationship

The interpretation of the statements made above in the section on two qualitative variables about the statistical association between them is not obvious. Recall that the two variables were the level of socialization of workers with their peers in a factory and their desire to stay or quit their job. We had found that the two variables were associated statistically. But there could be several possible interpretations of that statistical association.

**First interpretation:** We can interpret the statistical association to mean that a high level of socialization induces people to want to stay in that job. The explanation could be that the job is therefore more enjoyable, and people want to continue working there. In a way, the high level of socialization can be considered to be a cause for staying in that job, and inversely, a low level of socialization a reason to leave. So, we are now talking about more than a statistical association: we are talking about a **relationship** between variables. This situation can be represented by the diagram shown in Figure 8.5.



Figure 8.5

In symbolic terms, if we designate the level of socialization by  $X$ , and the desire to quit the job by  $Y$ , we could write:

$$X \Rightarrow Y$$

We could go a little further in that interpretation. If, in our theoretical framework, we had used the variable *Satisfaction with the job*, denoted by  $Z$ , as a general concept, and the level of socialization as one indicator of that concept, we could now conclude that the relationships can be illustrated by Figure 8.6.

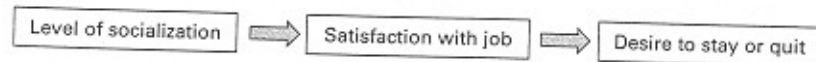


Figure 8.6

The following pattern illustrates the situation.

$$X \Rightarrow Z \Rightarrow Y$$

In other words, the level of socialization is used as an **explanatory** variable, to explain why people are more inclined to quit their jobs. Notice that this interpretation does not follow from the statistical analysis of the association between the two variables. This is clearly an interpretation, and it is not the only possible interpretation, as we will see in what follows.

**Second interpretation.** We could reverse the preceding interpretation and say that if individuals tend to quit their job (they may perhaps want a better salary, or a more challenging job), they will not invest a lot of energy in socializing with their peers, since they know they are going to quit soon. Here the model is reversed:

$$Y \Rightarrow X$$

In other words, the desire to quit the job is used to explain why people do not socialize a lot with their peers. This interpretation, like the previous one, does not follow automatically from the statistical association between the two variables. The statistical association allows such an interpretation, but it does not prove it.

**Third interpretation.** The results of the statistical analysis are consistent with yet another interpretation, which asserts that both the desire to quit and the lack of socialization are the result of a third variable, such as *Desire to get a better salary*. If people think that their present salary is too low, and that they can get a better salary if they find another job, they may plan to quit and also they may decide not to invest too much energy and time in socializing with their peers. The model proposed here for explaining the statistical association is the following.



**Fourth interpretation.** The last interpretation that we could propose is to consider both variables as indicators of the general concept *Satisfaction with job*. This concept could be measured by several indicators: level of socialization, intention to stay, satisfaction with the salary level, pleasant atmosphere at the office, relationship of support and cooperation with the management, etc. In this interpretation, the key concept is the global satisfaction with the job. When people are globally satisfied, they are more likely to socialize with their peers, to consider staying in this job for a long time, etc.

Sometimes the qualitative relationship between two correlated variables is said to be *spurious*. To say that a relationship is **spurious** means that there is no logical link between the two variables, and that the statistical association is misleading. Such statistical association is often due to a third variable, but the logics linking each of

the two correlated variable with the third one are completely unrelated. A classical example is that of height and salary. It could turn out that there is a statistical association between the height of an individual and his or her salary for a given sample. But if we break down the sample studied into men and women, we find that within each group there is no relationship. What happens is that on one hand men tend to be taller than women, and on the other hand in most societies the social structure favors men over women and the former end up tending to have higher salaries. The two kinds of associations (sex and height; gender and salary) follow logics that are totally unrelated to each other, hence our conclusion that the statistical association between height and salary is spurious. However, it is not always clear whether two sets of causal relationship are related or not, and one should be quite careful in interpreting a statistical association as spurious or as meaningful.

## Summary and Conclusions

### From Statistical Association to Relationship between Variables

The discussion above should help us understand better two distinct concepts, the concept of *statistical association* and the concept of *relationship between variables*.

**Statistical association** is something that can be observed objectively and measured, as we have seen in the examples above. Basically, it means that if you know the score of an individual on a variable  $X$  you can make a better guess of his or her score on another variable  $Y$  than if you did not know the score on  $X$ . The measure of statistical association depends on the level of measurement of the variables, which depends partly on the type of variables.

- For quantitative variables measured by a numerical scale, statistical association is called correlation. Two such quantitative variables are correlated when the values of one of them can be predicted with some precision from the values of the other variable. For linear correlation, the points representing the individuals are close to a straight line, which is called the regression line. If the association is strong, the points are very close to the line, the correlation coefficient  $r$  is close to 1 or  $-1$ , and the predictions based on the regression line involve a small error.
- For qualitative variables measured by a nominal scale, statistical association is analyzed with the help of a contingency table, also called a two-way table or a cross-tabulation. Statistical association means that individuals who are in a given category of the independent variable are more likely to be in a specific category of the dependent variable than in other categories. There are ways of measuring the strength of the association but they will not be discussed here.
- If one variable ( $X$ ) is quantitative (measured by a numerical scale) and the other one ( $Y$ ) qualitative (measured by a nominal scale), statistical association is studied by comparing the average scores on  $X$  across the various categories of  $Y$ .

This situation is summarized in Figure 8.7.

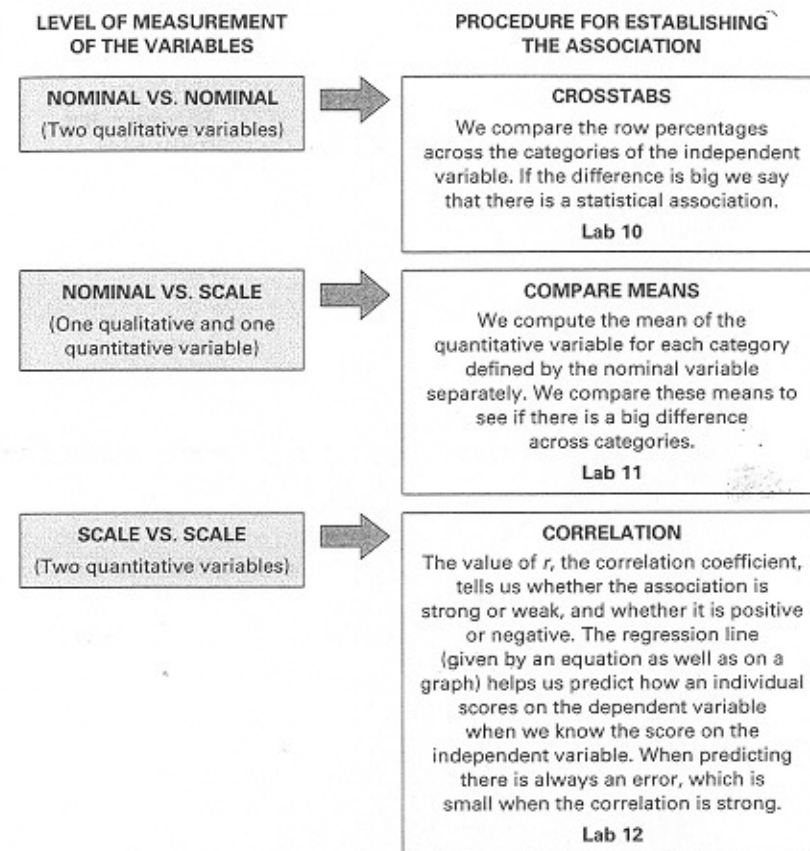


Figure 8.7 How to measure statistical association? It depends on the level of measurement of the variable

**Relationship between variables.** This notion is used to describe the *logical link* between variables. The independent variable could be a *cause* of the dependent variable, or an *explanatory factor* of the dependent variable; they could both be effects of some other variable; or they may be two indicators of a concept, or even two aspects of the same phenomenon. The notion of relationship between variables is a qualitative notion. It is a matter of interpretation, and it depends on the theoretical framework used in the research and on the research question or the research hypothesis. Statistical association should not be automatically interpreted as meaning a causal link.