Antonius Rachard. 2003. *Interpreting Quantitative Data with SPSS.* London: Sage.

8 STATISTICAL ASSOCIATION

The purpose of this chapter is to examine the basic meaning of statistical association with its important features (link, tendency, prediction, and strength), and then to see how statistical association is detected and measured depending on the level of measurement of the variables involved. The interpretation of statistical association as a qualitative relationship between the variables (explanation, possible causal factor, spurious association or other) is briefly discussed.

After studying this chapter, the student should know:

- the concept of statistical association and the fundamental aspects of a statistical association (link, tendency, prediction, strength);
- how to analyze association, depending on the measurement level of the variables;
- how to produce and read a two-way table (manually and with SPSS);
- how to produce and interpret a coefficient of correlation and a scatter plot;
- how to compare the mean of various subgroups on a variable;
- how to interpret a regression line, estimated scores, and errors in estimates;
- how to use the regression equation to predict a dependent variable;
- the difference between a statistical association and a relationship between variables;
- how to distinguish between the notions of explanation, causal factor, and spurious relationship.

The concept of statistical association is fundamental in research methodology. This concept allows us to formulate a clear notion of a *link* between variables when we notice that the scores of one individual on two different variables may somehow be related. But what do we mean by the word *related*? And how do we decide whether scores are related or not? Does it have to apply to every individual? Are there degrees in such relationships? What is the real meaning of statistical association? Does it mean that one factor is the cause of the other?

The notion of *statistical association* is quite abstract and it may be fuzzy for now, but we will gradually develop a detailed understanding of what it means.

Let us start with several examples.

- A teacher may notice that students who have good grades in mathematics tend to have good grades in physics as well.
- A doctor may notice that her female patients tend to be more resistant to certain kinds of infections than her male patients.
- A market study may demonstrate that people who like classical music tend to appreciate going to the opera more than those who do not like classical music do.

What do these statements exactly mean? Let us examine the first of our examples, which deals with the relationship between grades in mathematics and in physics. Suppose we have a class with the grades listed in Table 8.1.

Table 8.1

Student number	Grade in mathematics	Grade in physics
1	75	77
2	67	66
3	45	52
4	56	51
5	87	89
6	90	73
7	59	58
8	93	92
9	78	79
10	74	72
11	76	73
12	68	71
13	84	85
14	87	84
15	82	83
16	89	86
17	69	72
18	58	61
19	62	63
20	67	69
21	73	75

If we were to plot a scatter diagram of these grades in the two disciplines, we would get Figure 8.1.

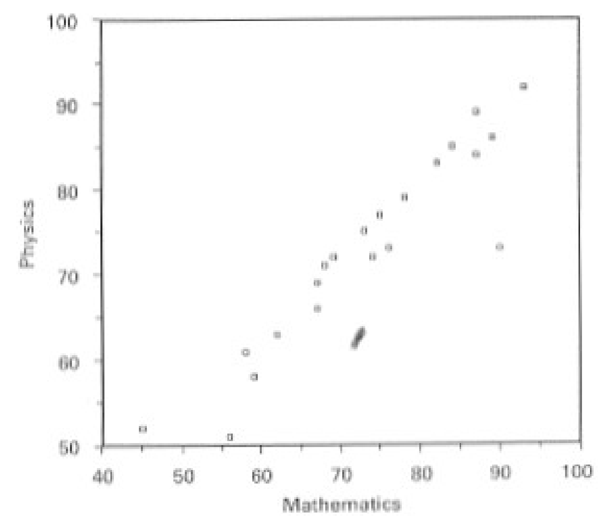


Figure 8.1 **Grades in mathematics and physics for a high school class**

Each dot represents one individual; the position of the dot with respect to the X-axis gives the grade of the individual in mathematics, and its position with respect to the Y-axis gives his or her grade in physics. Now we can identify several features in this diagram:

- When an individual scores low in mathematics, he/she tends to score low in physics as well.
- When he/she scores high in mathematics, he/she tends to score high in physics.
- Individuals whose score is close to the average in mathematics also tend to score close to the average in physics.
- The preceding remarks reflect a *tendency* and not a *rule*. You may have noticed that we always say that individuals who score in a certain way in mathematics *tend* to score in a certain way in physics. We can see that one individual does not fit the pattern outlined above, as this individual has a high grade in mathematics but a low grade in physics. This is why we talk about a *tendency* and not a *rule*.
- The notion of *prediction* is very important when we have a statistical association. If we know that somebody got a good grade in mathematics, we can *predict*, without knowing it, that his grade in physics is *likely* to be high. We see from the diagram above that we are right most of the time, but not all the time. Some individuals do not fit the pattern. This is why we use words like 'is likely to'. Predictions based on statistical association include a certain amount of error, in the sense that the predicted score differs from the real score by a certain amount, which is called the *error*. Such predictions also include a certain amount of risk, in the sense that there is a chance we are completely off track (as is the case if we tried to predict the grade in physics of the individual who got a good grade in math but a poor grade in physics).
- The notions of **dependent** and **independent variables** are used in this context. The *dependent variable* is what is to be explained, or what is to be predicted. The *independent variable* is the explanatory variable, or the variable used to make the prediction. In the example of the grades, the grade in mathematics is the independent variable and the grade in physics is the dependent variable. These two notions are not intrinsic to the variables, and the positions of dependent and independent variable could be interchanged, as we may want to see whether the grade in physics predicts the grade in mathematics with some accuracy.
- There are ways of measuring how *strong* an association is. The notion of *strength of an association* is related to that of prediction: if an association is strong, predictions based on it will tend to be good and will involve a small error. But if the association is weak, predictions based on it will often be way out … and involve large errors.
- The real concern here is to see whether there is some deep reason why people who perform well in mathematics also tend to perform well in physics. In some cases such a deep relationship exists, and in some others the statistical association is not indicative of a deep relation. Settling the issue of the existence of a relationship between variables is the real reason why we study statistical association. For the time being, let us remember that the existence of a statistical association is not a sufficient reason to say that there is a deep link between two variables.

The features outlined above express the essence of the notion of statistical association. But what if the variables are not quantitative? What does statistical association mean then? We will have to develop this notion separately for the various levels of measurements, and then draw some general conclusions. We will start by examining the case of two quantitative variables more closely.

The Case of Two Quantitative Variables

Let us suppose we have two quantitative variables, such as the grades of a class of students in mathematics and in physics in the example given above. We will denote the first one by X and the second one by Y. The grades of the various individuals in mathematics will be referred to as x_1, x_2, x_3, etc. and in physics as y_1, y_2, y_3, etc. When we want to talk about an individual in general, without saying which case this is, we will use the letter i. The situation is summarized in Table 8.2.

Table 8.2

Variable name	Symbol used	Entries are denoted by	General entry denoted by
Grade in Mathematics	X	x_1, x_2, x_3, etc	x_i
Grade in Physics	Y	y_1, y_2, y_3, etc.	y_i

Now we can start looking in more detail at the situation. Suppose the first student in the list has obtained 75 out of 100 in mathematics, and 77 out of 100 in physics, that is

$$x_1 = 75 \quad \text{and} \quad y_1 = 77.$$

This individual will be represented by the dot whose coordinates are (75, 77).

By looking at the scatter diagram shown in Figure 8.1, we can see a pattern. All the dots tend to fall on or near a straight line, called the **regression line**, shown in Figure 8.2.

This regression line represents the *trend* displayed by the dots. It can be described precisely by a mathematical equation (shown here at the top of the diagram). It can be used to **predict** the expected score in physics if the score of an individual in mathematics is known. On the diagram, you can see that somebody who scores 85 in mathematics is expected to score around 82 in physics: this is what the regression line suggests visually. If we want to calculate that predicted score more precisely, we could use the mathematical equation shown in the diagram, replacing x by the value 85. In this equation, y is the **predicted** value corresponding to a grade x in mathematics. This is what we get:

$$y = 11.523 + 0.83757x$$

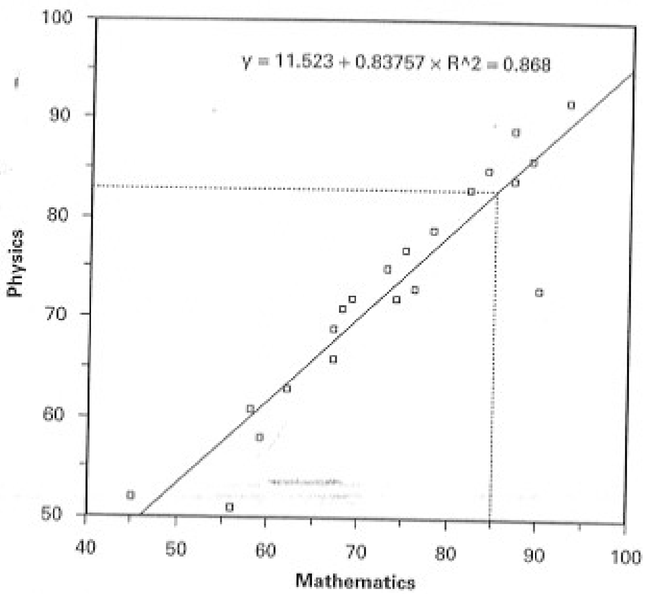


Figure 8.2 **Grades in mathematics and physics for a high school class**

If we replace x by 85 we get:

$$\text{predicted value of } y = 11.523 + 0.83757\ (85) = 82.71$$

or 83 if we round up. You will notice that this is the predicted value. It is the *expected* score of the individual. Thus, the regression line and its equation allow us to predict the scores in physics of an individual whose score in mathematics is known. Some individuals' real score will be slightly above or slightly below the expected value. In one of the cases shown in Figure 8.2, the expected score will be *very* different from the real score: this is the case of the individual represented by the dot on the lower right of the diagram.

But how good are these predictions generally? Can we measure how good they are? The answer is Yes. To understand it, consider the situation of one individual, illustrated by Figure 8.3.

If the individual is far away from the regression line, using the regression line for prediction will yield a large error. But if the individual is close to the regression line, the error in predicting his or her y-score will be small.

When we consider the whole population from the point of view of prediction, we get six types of situations shown in Figure 8.4, diagrams (a) to (f).

In diagram (a), the points that form the scatter diagram and that represent individuals are all found to be close to the regression line. In this case, when the

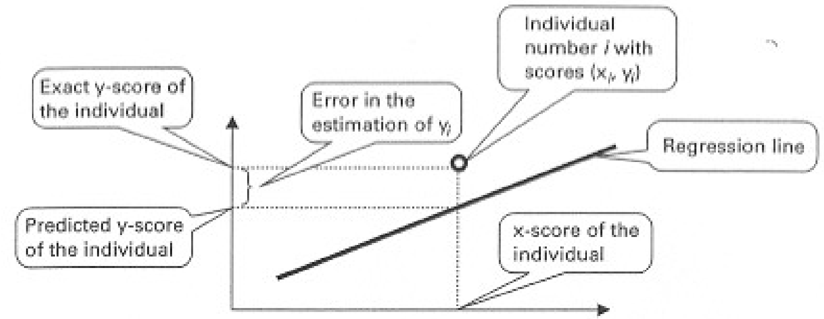


Figure 8.3

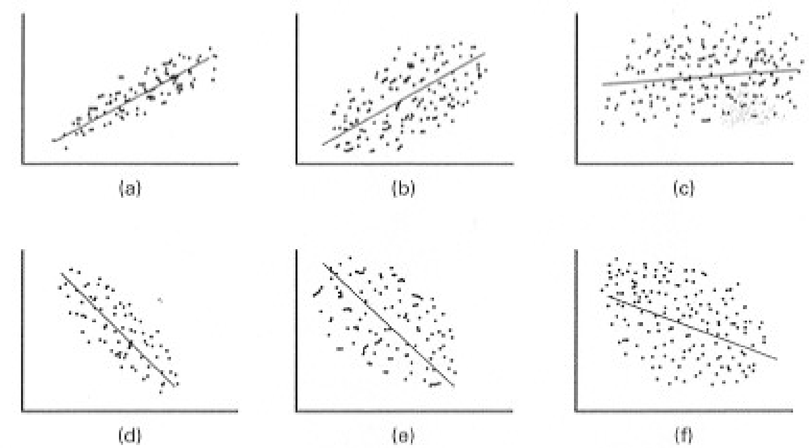


Figure 8.4

y-scores of individuals are predicted from their x-scores, the predictions tend to be generally good. We say in this case that **the correlation between the variable X and the variable Y is strong**. We used here the word *correlation* to refer to the statistical association. Indeed, correlation is the term to use when the variables are quantitative. Thus, the statistical association between quantitative variables is called a **correlation**.

In diagram (b), the points are not that close. We can still predict the y-score of an individual from his or her x-score, but the errors in prediction will tend to be larger than they were in diagram (a). In such a case, we say that the association between X and Y is not very strong.

In diagram (c), we see that the points are scattered far away from the regression line. People with high scores on the variable X do not tend to get high scores on Y: their scores on Y could be anywhere from low to high. In such cases, we say that **the correlation is weak** or even **null**.

The three remaining diagrams, (d), (e), and (f), are very similar to the preceding ones, with one difference that you may have noticed: as the x-scores increase, the y-scores tend to *decrease*. In such situations the **correlations are said to be negative.** They could be strong and negative, or weak and negative. The first correlations (a) to (c), in contrast, are said to be **positive.**

We have seen that some associations are weak (they yield poor predictions of the y-scores) and some are strong (they yield good predictions of the y-scores). In both cases they can be positive or negative. The next question now is to see **whether we can measure the strength of an association.**

There is indeed a mathematical formula that uses all the x- and y-values of the data to calculate the errors of prediction made on the basis of the regression line, and that comes up with a single number that summarizes it all. That number is called the **correlation coefficient**. It is obtained by the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x s_y}$$

where

x_i and y_i are the ith entry for X and Y respectively
$\bar{x}$ and $\bar{y}$ are the means of X and Y respectively, and
s_x and s_y are the standard deviations of X and Y respectively

This correlation coefficient is also referred to as the **Pearson product-moment correlation coefficient**. The values it produces range from −1 to +1. They can be interpreted as shown in Table 8.3.

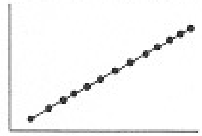
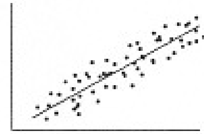
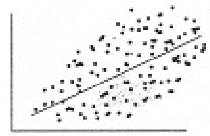
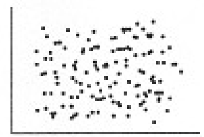
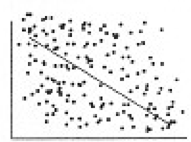
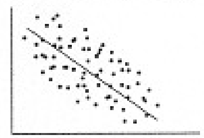
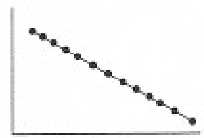
To illustrate the use of the correlation coefficient, we can consider the numerical example given above. The diagram indicated that $r^2 = 0.868$, which corresponds to $r = 0.93$ approximately, and that is a very strong correlation. In SPSS, a simple command allows you to get the program to compute r and r^2 for any two numerical variables. You will learn how to do that in Lab 12.

Warning: SPSS will compute the correlation coefficient even when the variables are not quantitative, provided the codes are numerical values. In such cases, the correlation coefficient is not meaningful. You should use the correlation coefficient and interpret it only when the variables are quantitative and measured by a numerical scale. The correlation coefficient can sometimes be used for quantitative variables measured at the ordinal level, but its interpretation is trickier and these situations should be avoided at this stage.

The Case of Two Qualitative Variables

How do we know that there is a statistical association between variables measured at the nominal level? The method of the correlation coefficient shown above does not apply. To illustrate the situation, we will take a concrete example and analyze it.

Table 8.3 **Meanings of the various values of the correlation coefficient**

Value of r	Value of r^2	Meaning	Scatter diagram illustrating it
$r = 1$	$r^2 = 1$	The correlation is perfect and positive. All the points fall exactly on the regression line.	
$r = 0.8$	$r^2 = 0.64$	The correlation is positive and strong. The points are fairly close to the regression line and the predictions based on it tend to be good. *As r decreases, the correlation is still positive but weaker, the points tend to be scattered away from the regression line and the predictions are increasingly poor.*	
$r = 0.3$	$r^2 = 0.09$	Very weak positive correlation. Poor prediction of y on the basis of knowing x.	
$r = 0$	$r^2 = 0$	The correlation is null. Knowing the value of x does not tell us anything about the likely value of y.	
$r = -0.3$	$r^2 = 0.09$	Very weak negative correlation. Poor prediction of y on the basis of knowing x. *As r takes larger negative values, the negative correlation gets stronger, the points tend to be closer to the regression line and the predictions are increasingly better.*	
$r = -0.8$	$r^2 = 0.64$	The correlation is negative and strong. The points are fairly close to the regression line and the predictions based on it tend to be good.	
$r = -1$	$r^2 = 1$	The correlation is perfect and negative. All the points fall exactly on the regression line.	

In a survey conducted in a large company, 300 employees were asked whether they are socializing with their peers at work at a high level or at a low level, and whether they were planning to look for another job. Their answers were compiled in Table 8.4. Every rectangle in the table is called a **cell**. The numbers in the cells refer to the frequency of each category, and are called **observed frequencies**.

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 *100.0%*	65 *100.0%*	300

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 × 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.

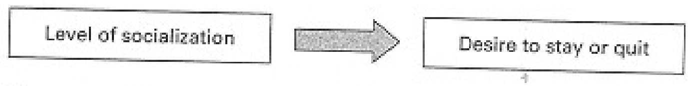


Figure 8.5

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.

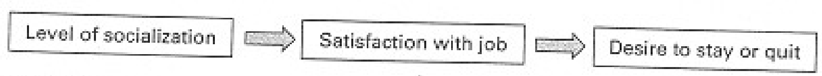


Figure 8.6

The following pattern illustrates the situation.

$$X \Rightarrow Z \Rightarrow Y$$

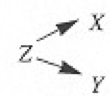
In other words, the level of socialization is used as an **explanatory** variable, to explain why people are more inclined to quit their jobs. Notice that this interpretation does not follow from the statistical analysis of the association between the two variables. This is clearly an interpretation,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

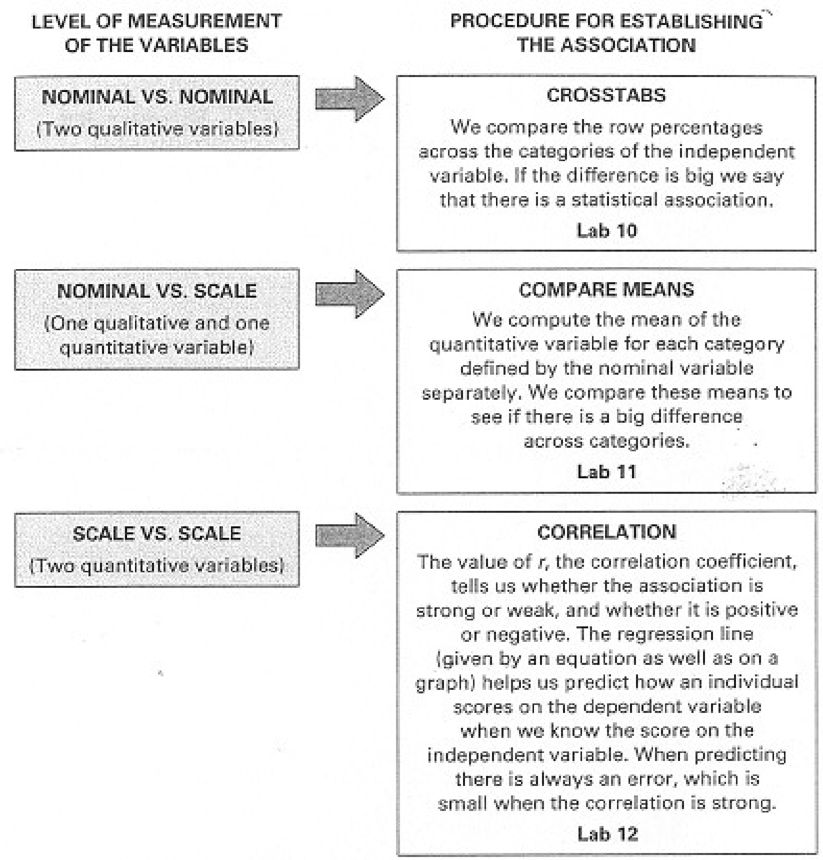


Figure 8.7 **How to measure statistical association? It depends on the level of measurement of the variable**

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.