

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.

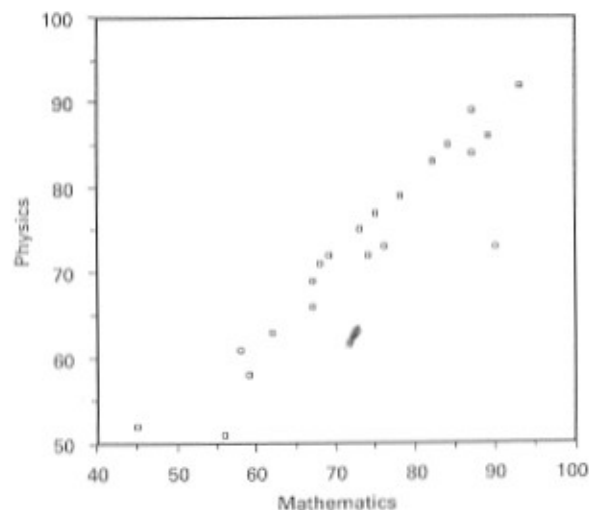


Figure 8.1 Grades in mathematics and physics for a high school class

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:

$$y = 11.523 + 0.83757x$$

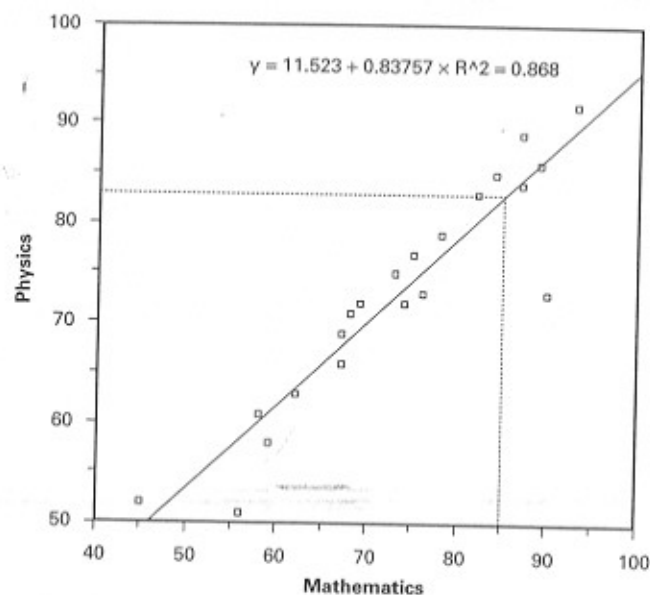


Figure 8.2 Grades in mathematics and physics for a high school class

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

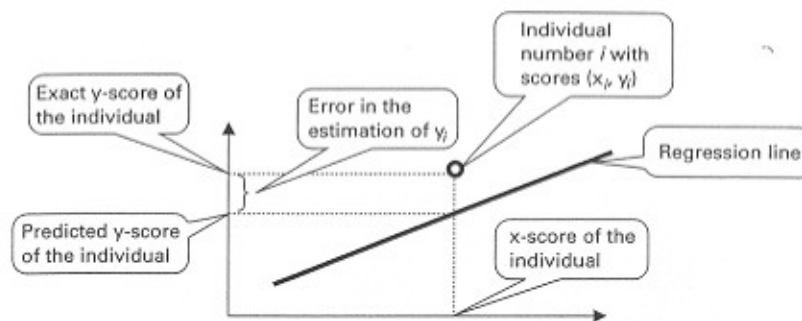


Figure 8.3

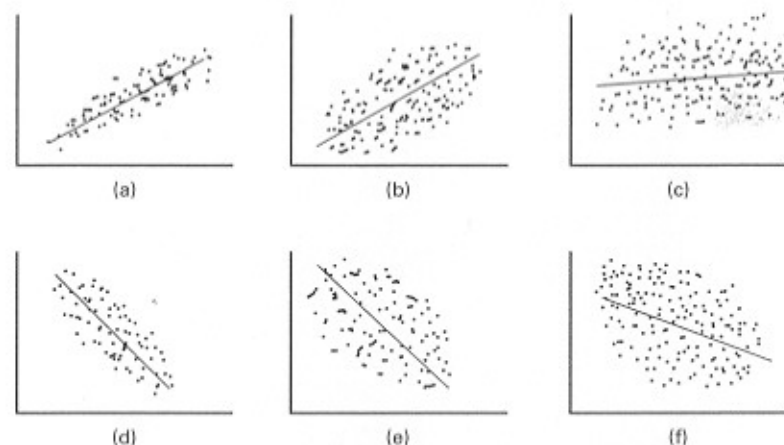


Figure 8.4

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 i th 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.	
		<i>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.</i>	
$r = 0.3$	$r^2 = 0.09$	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 .	
		<i>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.</i>	
$r = -0.8$	$r^2 = 0.64$	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**.