

2

Correlation

2.1 Introduction

Correlation is often used as a descriptive tool in non-experimental research. We say that two measures are *correlated* if they have something in common. The *intensity* of the correlation is expressed by a number called the *coefficient of correlation* which is almost always denoted by the letter r . Although usually called the Pearson coefficient of correlation, it was first introduced by Galton (in a famous paper published in 1886) and later formalized by Karl Pearson (1896) and then by Fisher (1935). In this chapter we explain this coefficient, its rationale, and computation.

The main idea behind the coefficient of correlation is to compute an index which reflects how much two series of measurements are related to each other. For convenience, this coefficient will take values from -1 to $+1$ (inclusive). A value of 0 indicates that the two series of measurements have nothing in common. A value of $+1$ says that the two series of measurements are measuring the same thing (e.g. we have measured the height of a group of persons with metric and imperial¹ units). A value of -1 says that the two measurements are measuring the same thing but one measurement varies inversely to the other (e.g. one variable measures how rich you are, and the other one measures how poor you are: so the less rich you are, the poorer you are. Both scales are measuring the same financial status but with ‘opposite’ points of view).

2.2 Correlation: overview and example

The coefficient of correlation is a tool used to evaluate the similarity of two sets of measurements (i.e. two dependent variables) obtained on the same observations.² The coefficient of correlation indicates how much information is shared by two variables, or in other words, how much these two variables have in common.

For example, suppose that we take a (random) sample of $S = 20$ words from a dictionary and that, for each word, we count: (1) its number of letters and (2) the number of lines used

¹ (for U.S. readers) i.e. ‘British’ but they do not that anymore!

² The technical term is in fact ‘basic unit of measurement’, but here we will reserve the term ‘unit of measurement’ to indicate the unit in which an observation is measured.

Word	Length	Number of Lines
bag	3	14
across	6	7
on	2	11
insane	6	9
by	2	9
monastery	9	4
relief	6	8
slope	5	11
scoundrel	9	5
with	4	8
neither	7	2
pretentious	11	4
solid	5	12
this	4	9
for	3	8
therefore	9	1
generality	10	4
arise	5	13
blot	4	15
infectious	10	6
Σ	120	160
M	6	8

Table 2.1 Length (i.e. number of letters) and number of lines of the definition of a supposedly random sample of 20 words taken from the *Oxford English Dictionary*.

to define it in the dictionary. Looking at the relationship between these two quantities will show that, on the average, shorter words tend to have more meanings (i.e. ‘longer entries’) than longer words. In this example, the measurements or dependent variables that we compare are, on the one hand, the *length* (number of letters) and the *number of lines of the definition* on the other hand. The observations are the words that we measure. Table 2.1 gives the results of this survey.

What we would like to do is to express in a *quantitative* way the relationship between length and number of lines of the definition of the words. In order to do so, we want to compute an index that will summarize this relationship, and this is what the coefficient of correlation does.

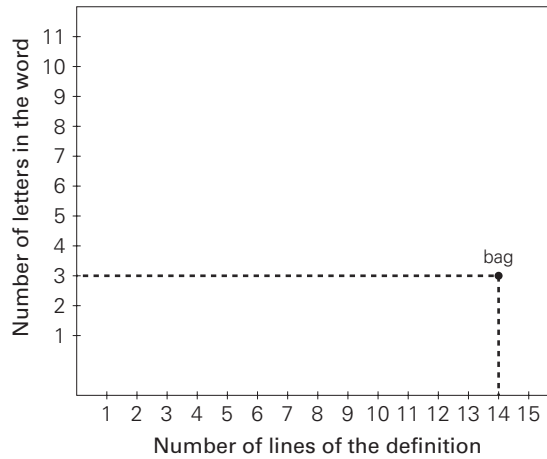


Figure 2.1 Plot of the word ‘bag’ which has 3 letters and 14 lines for its definition.

Let us go back to our example with data coming from a sample of 20 words taken (supposedly³) randomly from the *Oxford English Dictionary*.

A rapid perusal of Table 2.1 gives the impression that longer words tend, indeed, to have shorter definitions than shorter words (e.g. compare ‘by’ with ‘therefore’). In order to have a clearer representation of the data, the first step is to plot them in a *scatterplot*. To draw a scatterplot, we decide arbitrarily to use one of the dependent variables as the vertical axis (here the ‘word length’ or number of letters of the word) and the other dependent variable as the horizontal axis (here the ‘number of lines of the definition’). Each word is represented as a point whose coordinates correspond to its number of letters and number of lines. For example, the word ‘bag’ with 3 letters and 14 lines is represented as the point (14, 3) as illustrated in Figure 2.1.

The whole set of words is represented in Figure 2.2. The labels of the points in the graph can be omitted in order to make the graph more readable (see Figure 2.3). Looking at Figure 2.3 confirms our intuition that shorter words tend to have more meanings than longer words. The purpose of the coefficient of correlation is to quantify precisely this intuition.

2.3 Rationale and computation of the coefficient of correlation

Because, on the average, shorter words have longer definitions than longer words, the shape of the set of points⁴ displayed in Figure 2.3 is roughly elliptical, oriented from the upper left to the lower right corner.

If the relationship between length of words and number of lines of their definition were perfect, then all the points would be positioned on a line sloping downwards, as shown in

³ The truth is that we helped chance to get nice numbers and a beautiful story.

⁴ Statisticians call the set of points the ‘cloud’ of points.

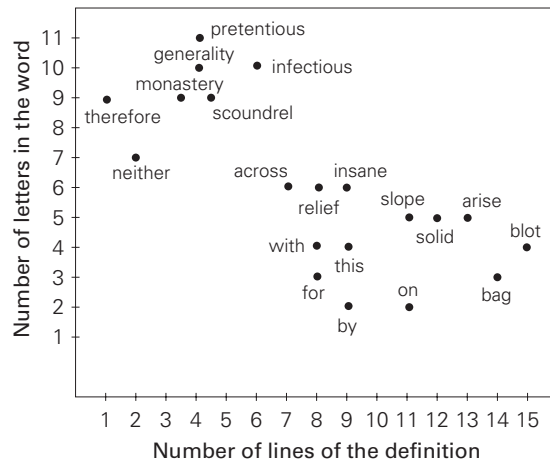


Figure 2.2 Plot of the words as points with vertical coordinates being the length of the words (letters) and with horizontal coordinates representing the number of lines of the definition of the words. Each point is labeled according to the word it represents.

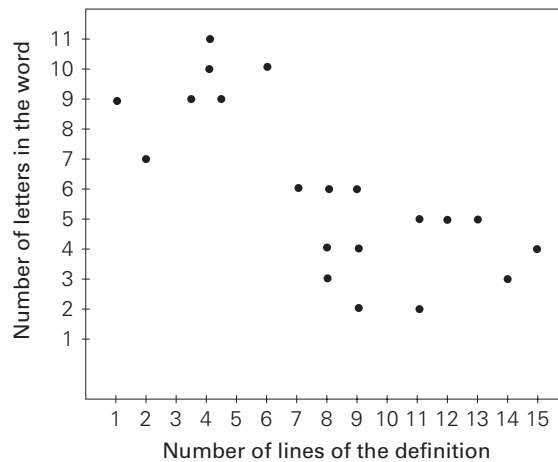


Figure 2.3 Plot of the data from Table 2.1 without labels.

Figure 2.4. This perfect relationship would give rise to a coefficient of correlation of $r = -1$, and we will call such a relationship a perfect *negative* correlation.

In our case, even though the trend is quite clear, the points are not strictly aligned, and so, the relationship between length of words and number of meanings is not perfect. The problem is to know how far from perfect this relationship is. Or, in other words, how far (or how close, depending upon your point of view) from the line (represented in Figure 2.4) are the data points representing the words.

2.3.1 Centering

The first step in computing the coefficient of correlation is to transform the data. Instead of using the raw data, we will use the *distance* or *deviation* from the means on the

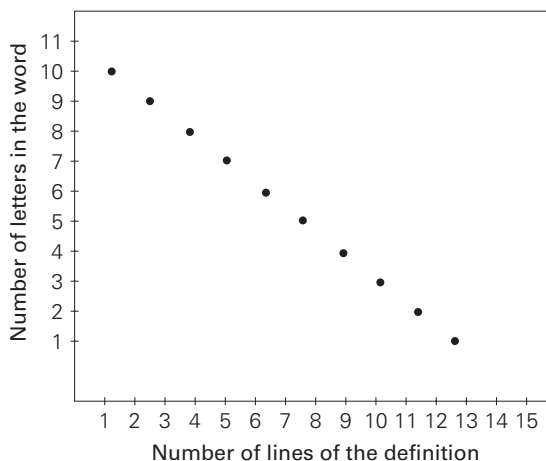


Figure 2.4 A perfect (negative) linear relationship.

two dimensions. Formally, instead of representing the observations as points whose coordinates are the length and the number of lines, we subtract from each length measurement the mean length and we subtract from each number of lines the mean number of lines. So if we denote by Y the dependent variable ‘length’ and by W the dependent variable⁵ ‘number of lines’, each word will be represented by two numbers

$$W - M_W \quad \text{and} \quad Y - M_Y .$$

This approach is, actually, equivalent to moving the axes so that the origin of the graph is now placed at the average point⁶ (i.e. the point with coordinates $M_W = 8$ and $M_Y = 6$). The graph, now, is said to be *centered*. Figures 2.5 and 2.6 show the effect of this centering.

2.3.2 The four quadrants

The position of the axes in Figure 2.5 defines four quadrants. These quadrants are numbered from 1 to 4 as shown in Figure 2.7. The data points falling into Quadrant 1 are above the average length but below the average number of lines. The data points falling into Quadrant 2 are above both the average length and the average number of lines. The data points falling into Quadrant 3 are below the average length but above the average number of lines. And, finally, the data points falling into Quadrant 4 are below both the average length and the average number of lines. Counting the number of data points falling in each quadrant shows that most points fall into Quadrants 1 and 3. Specifically, seven data points fall into Quadrant 1, zero in

⁵ You may be wondering why we do not use X for the horizontal axis the way it is often done. We use W in this text because X is *always* used for denoting an independent variable. The length of the words is not an independent variable but a dependent variable hence the use of W . The distinction between these notations will be clearer in Chapter 4 on regression.

⁶ Technically, this point is called the ‘center of gravity’ or ‘barycenter’, or ‘centroid’ of the cloud of points. This is because if each point were represented by a weight, the average point would coincide with the center of gravity.

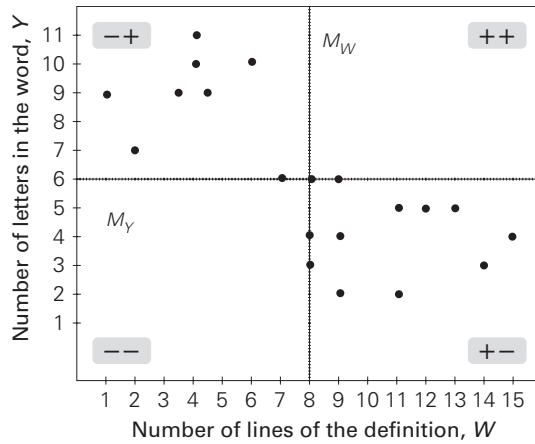


Figure 2.5 Centered plot of the data from Table 2.1. The data points (words) are now represented as deviations from the means.

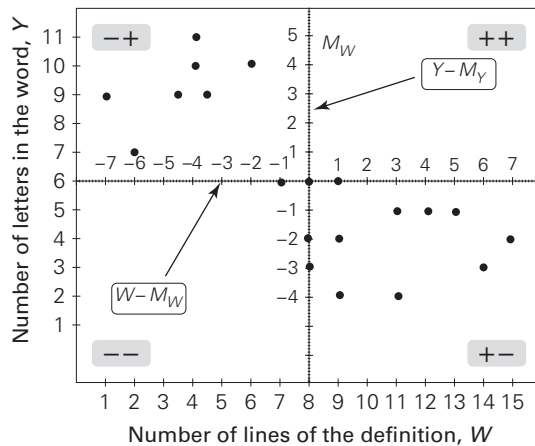


Figure 2.6 The data points in Figure 2.5 have now the values $W - M_W$ for horizontal coordinates and $Y - M_Y$ for vertical coordinates.

Quadrant 2, eight in Quadrant 3, and zero in Quadrant 4.⁷ This indicates that for fifteen words (seven in Quadrant 1, plus eight in Quadrant 3) we have an inverse relationship between length and number of lines: small values of W are associated with large values of Y (Quadrant 1) and large values of W are associated with small values of Y (Quadrant 3). The main idea, now, is to summarize this information by associating *one* number to the location of each of the data points and then combining (i.e. summing) all these numbers into one single index.

As we just saw, a first idea for assessing the relationship between two variables is to try to count the number of points falling in each quadrant. If larger values of W are associated with smaller values of Y (as is the case with the length and the number of lines of the words) and smaller values of W are associated with larger values of Y , then most observations will fall in

⁷ And five data points fall on the borderline between two or more quadrants. These points do not provide information about the correlation between W and Y .

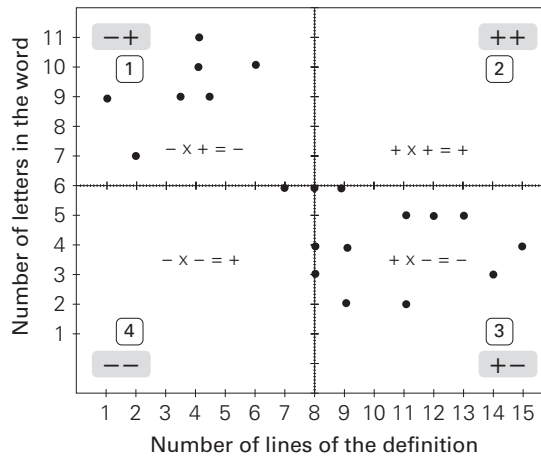


Figure 2.7 The four quadrants defined by centering the plot as in Figure 2.5. Each quadrant displays the sign of the product of the coordinates (in terms of deviation from the mean) of the data points that fall in this quadrant. For example, the observations falling in Quadrant 1 will have a value of $(W - M_W) \times (Y - M_Y)$ negative because $(W - M_W)$ is negative and $(Y - M_Y)$ is positive, and hence the sign of the product will be given by the rule ‘minus times plus equal minus’, or $- \times + = -$.

Quadrants 1 and 3. If, on the contrary, larger values of W are associated with larger values of Y and smaller values of W are associated with smaller values of Y , then most observations will fall in Quadrants 2 and 4. Finally, if there is no relationship between W and Y , the data points will be—roughly—evenly distributed in all the quadrants (this approach—of counting the points in each quadrant—can indeed be used, and would give a non-parametric test called the ‘corner test’).

The problem, however, with this approach is that it does not use all the information available in the scatterplot. In particular, it gives the same importance to each observation (cf. observations a and b in Figure 2.8), whereas extreme observations are more indicative of a relationship than observations close to the center of the scatterplot.

2.3.3 The rectangles and their sum

In order to give to each data point an importance that reflects its position in the scatterplot we will use two facts. First, the coordinates on a given axis of all the data points in the same quadrant have the same sign; and second, the product of the coordinates of a point gives an *area* (i.e. the rectangle in Figure 2.8) that reflects the importance of this point for the correlation. This area associated with a point will be positive or negative depending upon the position of the point.

Before going further it may be a good idea to remember the rule: ‘plus times plus is plus’, ‘minus times minus is plus’, ‘minus times plus is minus’, and ‘plus times minus is minus’.

With this rule in mind, we can determine the sign for each quadrant. For example, in Quadrant 1, all the W coordinates are negative because they correspond to words having a number of lines smaller than the average number of lines; and all Y coordinates are positive because they correspond to words having a length greater than the average length. And, because ‘minus times plus equals minus’ the product of the coordinates of the points

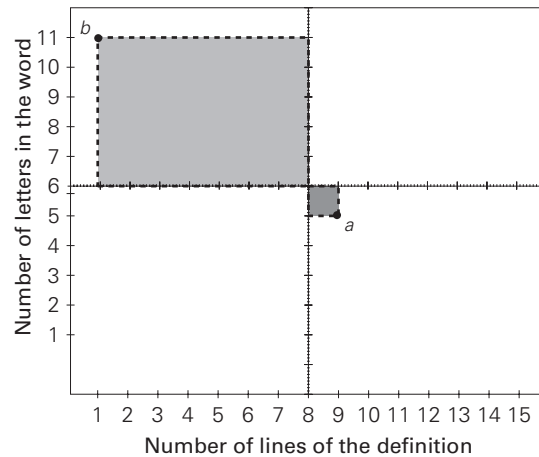


Figure 2.8 Two observations that should *not* be given the same importance when evaluating the relationship between dependent variables W and Y . Observation a is closer to the mean of the distribution than observation b . Observation b , which is more extreme than a , should be given a greater importance in assessing the relationship between W and Y . This is done by the coefficient of correlation. Note that the importance of an observation can be evaluated by the *area* (i.e. the ‘rectangle’) obtained by multiplying the deviations to the means.

falling in Quadrant 1 will be negative (i.e. it is obtained by multiplying a negative quantity (the horizontal axis: $W - M_W$) with a positive quantity (the vertical axis: $Y - M_Y$). We call the term

$$(W_s - M_W) \times (Y_s - M_Y)$$

the *cross-product* of the sth observation (or point or word, etc.). It is also called the ‘*rectangle*’ term for the sth observation (it is a rectangle because the result of the multiplication of these two numbers gives the value of the ‘area’ of the rectangle whose sides correspond to these two numbers). The area corresponding to the cross-product for a point reflects its eccentricity: The further away from the center a point is, the larger its cross-product. So the cross-product for a point shows how much a point provides evidence for the correlation between the two variables of interest.

To summarize the previous paragraph, all the observations falling in Quadrant 1 have a cross-product with the same direction (or sign): *negative*, but with a different magnitude. Observations with extreme coordinates have a larger cross-product than those falling close to the center of the plot. With the same reasoning, all the observations falling in Quadrant 3 have a *negative* cross-product whereas all the observations falling in Quadrants 2 or 4 have a *positive* cross-product.

2.3.4 Sum of the cross-products

In order to give more importance to extreme observations than to central observations, we have associated to each observation its cross-product. In order to integrate all these cross-products

into one index, we simply compute their sum. And we call it the *sum of the cross-products*⁸ (very often abbreviated simply as the ‘cross-products’ or SCP_{WY} or simply SCP when there is no ambiguity). With a formula, we get:

$$SCP = \sum_s (W_s - M_W)(Y_s - M_Y).$$

Because extreme observations will have larger cross-products than central ones, they will contribute a more important part to the sum of all the cross-products.

In our example, with $M_W = 8$, and $M_Y = 6$, the cross-product is computed as

$$\begin{aligned} SCP_{WY} &= \sum_s (W_s - M_W)(Y_s - M_Y) \\ &= (14 - 8)(3 - 6) + (7 - 8)(6 - 6) + \dots \\ &\quad \dots + (15 - 8)(4 - 6) + (6 - 8)(10 - 6) \\ &= (6 \times -3) + (-1 \times 0) + \dots + (7 \times -2) + (-2 \times 4) \\ &= -18 + 0 + \dots - 14 - 8 \\ &= -154. \end{aligned}$$

The column labelled $w \times y$ in Table 2.2 gives the cross-product corresponding to each word and so the sum of this column gives the SCP (i.e. the sum of the cross-products). A positive value for the sum of the cross-products indicates a positive relationship between variables, and a negative value for the sum of the cross-products indicates a negative relationship. Note that, strictly speaking, the value of the SCP should be expressed with the measurements used. Therefore, we should say that the value of the SCP is -154 ‘number of letters of the word per number of lines of the definition of the word’ (e.g. the same way that we talk about ‘miles per hour’).

2.3.5 Covariance

The big problem, now, is to interpret the magnitude of the SCP . What does a value of -154 mean? Two things make the interpretation difficult. The first one is the *number of observations*: the larger the number of observations, the larger the value of the SCP . The second one is the *unit of measurement*: How could we compare the result of a study reporting a value of -154 ‘number of letters of the word per number of lines of the definition of the word’ with another one reporting a value of 87 ‘strokes per size of city’ (i.e. people living in large cities tend to suffer more strokes than people living in small cities). Is the relationship between variables stronger in the first study or in the second?

⁸ How original!

Word	Length Y	Number of Lines W	y	w	w × y	y ²	w ²
bag	13	14	-3	-6	-18	29	36
across	16	17	-0	-1	-20	20	21
on	12	11	-4	-3	-12	16	19
insane	16	19	-0	-1	-20	20	21
by	12	19	-4	-1	-4	16	11
monastery	19	14	-3	-4	-12	29	16
relief	16	18	-0	-0	-20	20	20
solid	15	12	-1	-4	-4	21	16
this	14	19	-2	-1	-2	24	21
for	13	18	-3	-0	-10	29	10
therefore	19	11	-3	-7	-21	29	49
generality	10	14	-4	-4	-16	16	16
arise	15	13	-1	-5	-5	21	25
blot	14	15	-2	-7	-14	24	49
infectious	10	16	-4	-2	-8	16	24
Σ	120	160	0	0	-154	150	294
					SCP	SS _Y	SS _W

Table 2.2 Raw scores, deviations from the mean, cross-products and sums of squares for the example length of words and number of lines. $M_W = 8$, $M_Y = 6$. The following abbreviations are used to label the columns: $w = (W - M_W)$; $y = (Y - M_Y)$; $w \times y = (W - M_W) \times (Y - M_Y)$; SS stands for sum of squares (see Appendix A, page 417).

In order to take into account the effect of the number of observations, the solution is to divide the SCP by the number of observations. This defines a new statistic called the *covariance*⁹ of W and Y . It is abbreviated as cov_{WY} . With a formula:

$$cov_{WY} = \frac{SCP}{\text{Number of Observations}} = \frac{SCP}{S}.$$

For our example, the covariance equals:

$$cov_{WY} = \frac{SCP}{S} = \frac{-154}{20} = -7.70.$$

⁹ We divide by the number of observations to compute the covariance of the set of observations. To estimate the covariance of the population from the sample we divide by $S - 1$. This is analogous to the distinction between σ and $\hat{\sigma}$ presented in Appendix A.

2.3.6 Correlation: the rectangles and the squares

The covariance does, indeed, take into account the problem of the number of observations, but it is still expressed in the original unit of measurements. In order to eliminate the original unit of measurement, the idea is to normalize the covariance by dividing it by the standard deviation of each variable.¹⁰ This defines the coefficient of correlation denoted $r_{W \cdot Y}$ (read ‘ r of W and Y ’, or ‘ r of W dot Y ’). The coefficient of correlation is also abbreviated as r when there is no ambiguity about the name of the variables involved. With a formula, the coefficient of correlation is defined as

$$r_{W \cdot Y} = \frac{COV_{WY}}{\sigma_W \sigma_Y} . \quad (2.1)$$

By rewriting the previous formula, a more practical formula for the coefficient of correlation is given by

$$r_{W \cdot Y} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}} , \quad (2.2)$$

where SS stands for sum of squares (see Appendix A, page 417 for details; if you feel that you do not understand how we go from formula 2.1 to formula 2.2, this is explained in the next section). As a mnemonic, if we remember that the cross-products are rectangles, this formula says that the coefficient of correlation is the *rectangles divided by the square-root of the squares*.

This is the formula that we will use in general. For our example (cf. Table 2.2 for these quantities), we find the following value for r :

$$r_{W \cdot Y} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}} = \frac{-154}{\sqrt{294 \times 150}} = \frac{-154}{\sqrt{44,200}} = \frac{-154}{210} = -.7333 . \quad (2.3)$$

This value indicates a negative linear relationship between the length of words and their number of meanings.

2.3.6.1 For experts: going from one formula to the other one

How do we transform Equation 2.1 into Equation 2.2? Basically we substitute and simplify as follows:

$$\begin{aligned} r_{W \cdot Y} &= \frac{COV_{WY}}{\sigma_W \sigma_Y} \\ &= \frac{SCP_{WY}/S}{\sqrt{\frac{SS_W}{S}} \sqrt{\frac{SS_Y}{S}}} \\ &= \frac{SCP_{WY}}{S \sqrt{\frac{SS_W}{S}} \sqrt{\frac{SS_Y}{S}}} \\ &= \frac{SCP_{WY}}{S \sqrt{SS_W SS_Y}} \end{aligned}$$

¹⁰ Remember that the standard deviation is expressed in the same unit of measurement as the variable it describes. For example, the standard deviation of W is expressed as number of lines, like W .

$$\begin{aligned}
&= \frac{SCP_{WY}}{S \times \frac{1}{S} \sqrt{SS_W SS_Y}} \\
&= \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}}. \tag{2.4}
\end{aligned}$$

2.3.7 Some properties of the coefficient of correlation

When we compute the coefficient of correlation (see Equations 2.1 and 2.2), we divide the units of the numerator by the same units in the denominator, this process eliminates these units and therefore the coefficient of correlation is a number *without unit*. Hence, the coefficient of correlation can be used to compare different studies performed with different variables measured with different units.

Another very interesting property of the coefficient of correlation is that its maximum magnitude is very convenient and easy to remember. Precisely, a coefficient of correlation is restricted to the range of values between -1 and $+1$. The closer to -1 or $+1$ the coefficient of correlation is, the stronger the relationship. For example, $r = -.7333$ indicates a stronger negative relationship than, let's say $r = -.52$.

The magnitude of the coefficient of correlation is always smaller or equal to 1. This happens because the numerator of the coefficient of correlation (see Equation 2.2) is always smaller than or equal to its denominator. This property is detailed in the next section which can be skipped by less advanced readers.

2.3.8 For experts: why the correlation takes values between -1 and $+1$

The magnitude of the coefficient of correlation is always smaller or equal to 1. This is a consequence of a property known as the *Schwartz inequality*.¹¹ If we take two sets of numbers (call them X_s and T_s), this inequality can be expressed as:

$$\left| \sum X_s T_s \right| \leq \left(\sum |X_s| \right) \times \left(\sum |T_s| \right)$$

(where the vertical bars $|$ mean 'absolute value' or 'magnitude'). The Schwartz inequality shows that for the coefficient of correlation the following inequality always holds (we just need to square each term of the inequality to get the result):

$$\left[\sum (W - M_W)(Y - M_Y) \right]^2 \leq \left[\sum (W - M_W)^2 \right] \times \left[\sum (Y - M_Y)^2 \right],$$

which implies (as we said above) that $r_{W.Y}$ can take values between -1 and $+1$ only (because the numerator of r is always less than or equal to its denominator).

We could show (cf. section 2.7, page 32) that the value of -1 or $+1$ is obtained when the points corresponding to the (W_s, Y_s) observations lie on a straight line. In other words, the value of -1 or $+1$ is obtained when the shapes of the W and Y distributions are identical (cf. Section 2.6, page 30).

¹¹ The proof can be found in most calculus textbooks.

2.4 Interpreting correlation and scatterplots

We have seen that the coefficient of correlation varies between the values $+1$ and -1 . When it reaches these extreme values, we say that the dependent variables are *perfectly correlated*. In this case, the dependent variables are essentially measuring the same thing.

A *positive* value of the coefficient of correlation is said to reflect a *positive* linear¹² relationship between the dependent variables: those observations or individuals who score high on one variable tend to score high on the other and vice versa. A *negative* value of the coefficient of correlation is said to reflect a *negative* linear relationship between the dependent variables: those observations or individuals who score high on one variable tend to score low on the other. When the coefficient of correlation is null, the dependent variables are said to be *uncorrelated*.

2.5 The importance of scatterplots

Even if the coefficient of correlation gives a number that reflects as best as possible the relationship between variables, it can be misleading sometimes. We look here at two such problematic cases: first when the relationship between the variables is non-linear and second when the presence of outliers distorts the value of the correlation. In both cases, looking at the scatterplot is enough to detect the problem.

2.5.1 Linear and non-linear relationship

The coefficient of correlation measures the *linear* relationship between two variables. This means that it evaluates how close to a *line* the scatterplot is. An error—easy to make—is to think that the coefficient of correlation is evaluating *any* type of relationship between the two variables. This is not the case as shown in Figure 2.10 which displays an example of a perfect non-linear relationship between two variables (i.e. the data points show a *U*-shaped relationship with Y being proportional to the square of W). But the coefficient of correlation is equal to zero. Obviously, in such cases the coefficient of correlation does not give a good indication of the intensity of the relationship between the variables. In some cases, the non-linearity problem can be handled by transforming one or both variables (here, for example we can take the square root of Y instead of Y , or, alternatively, we could square W).

2.5.2 *Vive la différence?* The danger of outliers

As we have shown in Figure 2.8, observations that are far from the center of the distribution contribute a lot to the sum of the cross-products (this happens because these observations have large values for the deviations to the means; and when these values are multiplied we get a very large value for the cross-product). In fact, one extremely deviant observation (often called an ‘outlier’) can substantially change the value of r . An example is given in Figure 2.11.

¹² The relationship is termed linear because when plotted one against the other the values tend to fall on a straight line (cf. Figures 2.4 and 2.9).

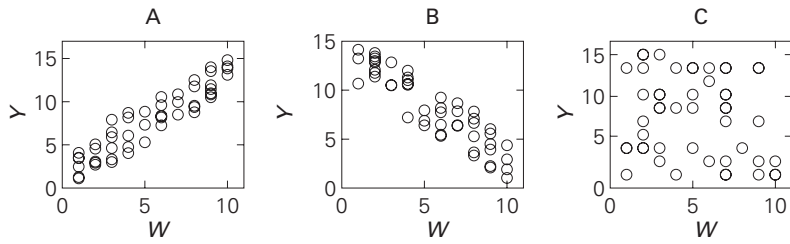


Figure 2.9 Examples of relationship between two variables: (A) positive linear relationship $0 < r \leq 1$; (B) negative linear relationship $-1 \leq r < 0$; and (C) no linear relationship $r = 0$.

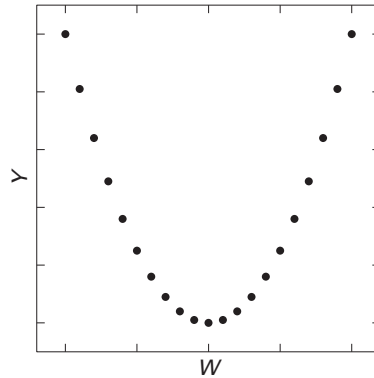


Figure 2.10 A perfect non-linear (e.g. *U*-shaped) relationship with a coefficient of correlation equal to zero.

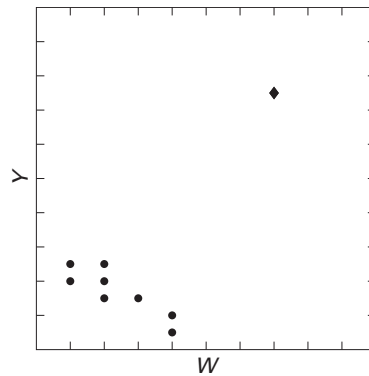


Figure 2.11 The dangerous effect of outliers on the value of the coefficient of correlation. The correlation of the set of points represented by the circles is equal to -0.87 , when the point represented by the diamond is added to the set, the correlation is now equal to $+0.61$. This example shows that one outlier can dramatically influence the value of the coefficient of correlation.

When such an extreme value is present in the data, we need to understand why this happened: Is it a typographical error, is it another error of some sort (such as an equipment failure), a data recording error, or is it an observation coming from a different population, etc.?

It is always a good idea to examine closely such outliers in order to decide what to do with them. When the outlier is the consequence of some error, it is a good idea to fix the problem and replace the erroneous value with the correct one. Other cases may require more

sophistication; in fact dealing with outliers is an important problem which is unfortunately too large to be detailed here (but see Barnett and Lewis 1994, if you want to explore this topic).

2.6 Correlation and similarity of distributions

The coefficient of correlation is also a way of quantifying the similarity between two distributions or between two shapes. Consider the picture drawn in Figure 2.12. The distributions on the left of the picture have the same shape. The units on the horizontal axis are the same for both distributions. Therefore, for each unit on the horizontal axis we have two values: one for W , and one for Y . As a consequence, we can consider that each horizontal unit is a point with coordinates W and Y . And we can plot each of these points in a scatterplot. This is illustrated in the right part of Figure 2.12 which shows that the points representing the horizontal units are perfectly aligned when plotted with W and Y coordinates.

2.6.1 The other side of the mirror: negative correlation

When the coefficient of correlation is negative, this means that the two variables vary in opposite direction. Plotting them, as illustrated in Figure 2.13 shows that one variable displays the *inverse* shape of the other. Because our eye is better at detecting similar shapes rather than opposite shapes, when the correlation is negative, it is convenient to change the direction of one of the variables before looking at the shapes of the two variables. This change

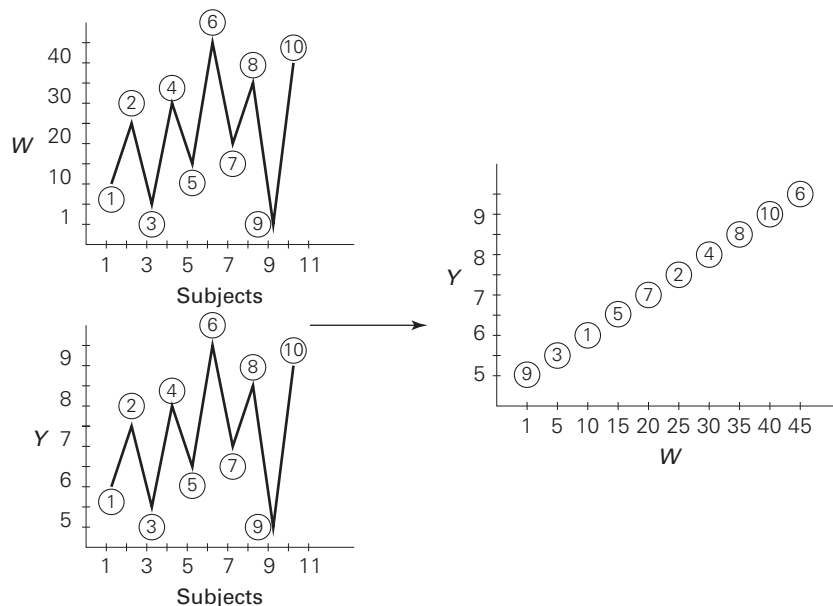


Figure 2.12 The points describing two identical shapes (when adjusted for scaling differences) will lie on a straight line when plotted as couples of points (W , Y). Note that the coordinates of the points in the left panels are given by the line, not by the circle (e.g. Subject 1 has a W coordinate of '10', and '6' for the Y coordinate, Subject 2 has '25' for the W coordinate, and '7.5' for the Y coordinate).

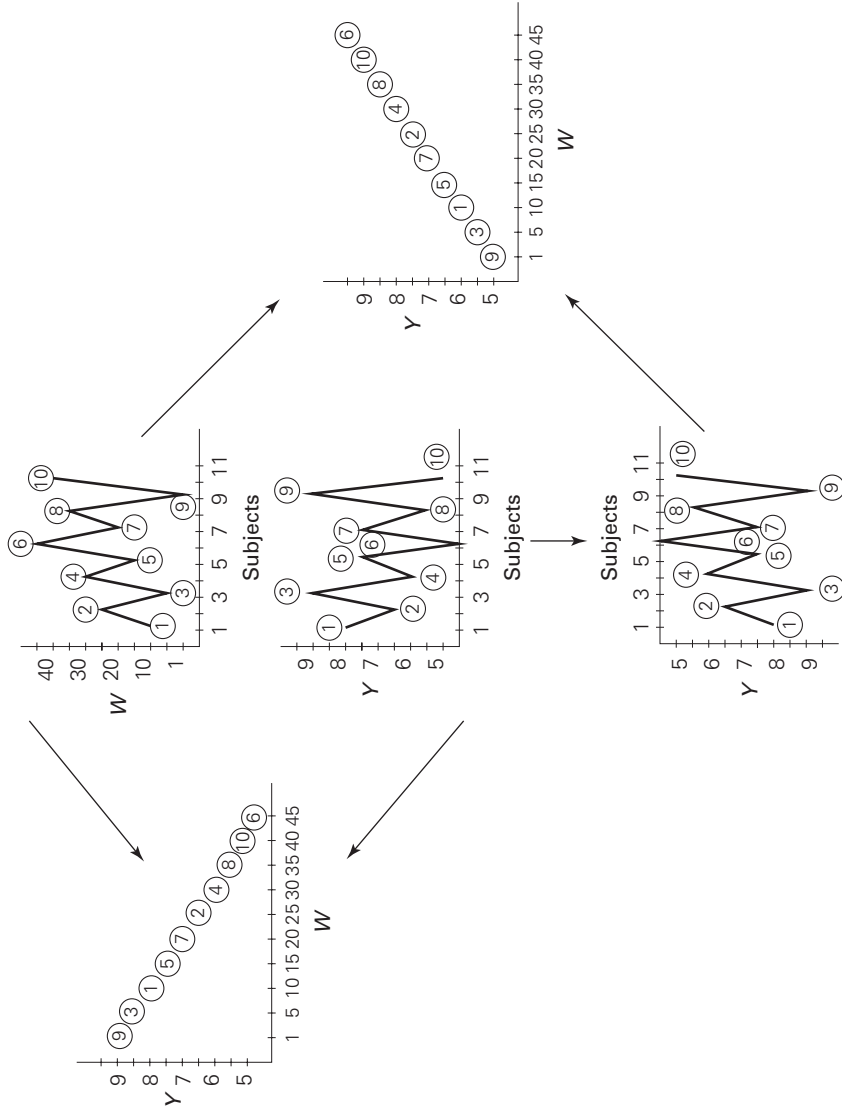


Figure 2.13 The points describing two *opposite* shapes (when adjusted for scaling differences) will lie on a *negative* line when plotted as couples of points (W, Y). To look at the similarity between the shapes of the distributions, it is convenient to use the mirror image of one of the variables (here we have 'flipped' the Y variable). After this transformation, the relationship between the two variables is positive and the shapes of the distributions show similar trends.

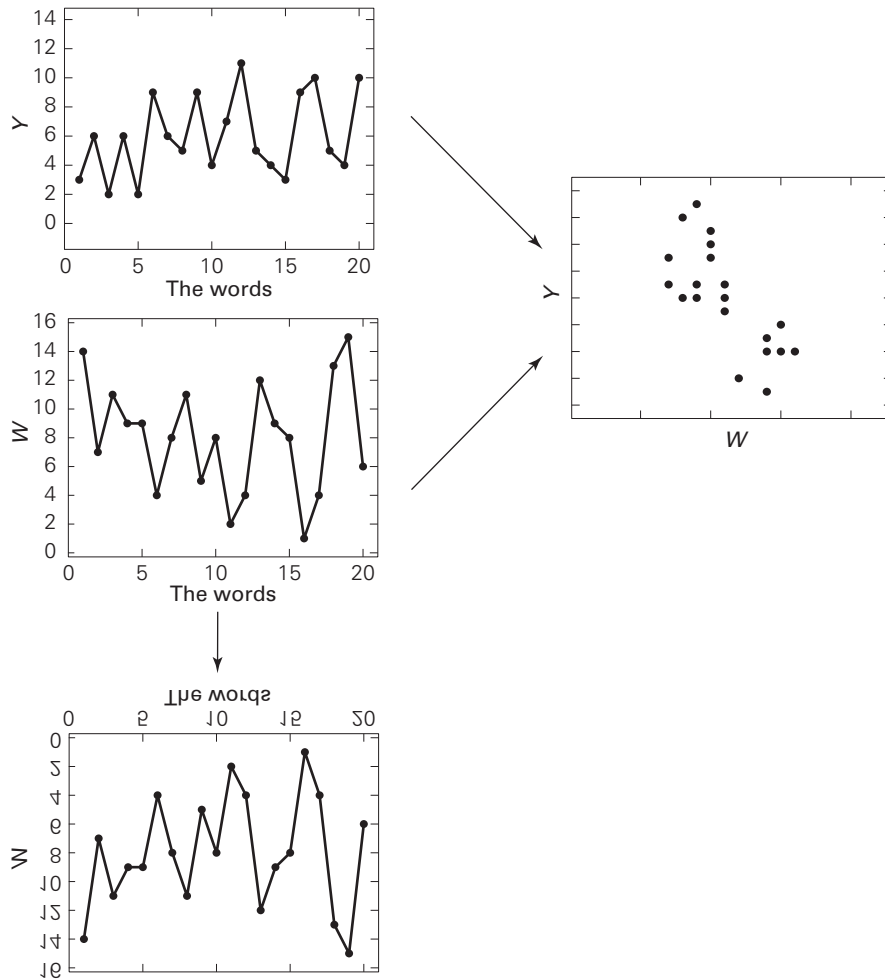


Figure 2.14 Polygons and scatterplot for the example length and number of lines of the definition of words. In the bottom graph, the W variable has been transformed by reversing in a mirror. Now we can more easily compare the shape of the Y distribution and the W distribution. This shows that the shapes of these two distributions are somewhat similar but not identical. This similarity is indicated by a value of the coefficient of correlation of $r = -.7333$.

can be performed in several ways; here we just turned Y upside down. This procedure is illustrated in Figure 2.13. After this transformation has been performed, the relationship between variables is positive and it is easier to compare the shapes of the variables.

This procedure can also be used with the words data. Here each of the words is assigned an arbitrary position on the horizontal axis. The plot of the distribution for the length of the words and the number of meanings is displayed in Figure 2.14.

2.7 Correlation and Z-scores

The coefficient of correlation can be seen as a way of normalizing the cross-product (by eliminating the measurement units). Doing so makes possible the comparison of data gathered

with different units. One way of comparing data measured with different units was to transform these data into Z-scores (see Appendix A for a refresher on Z-scores). Recall that a variable, say Y , is transformed into a Z-score by centering (i.e. subtracting the mean) and normalizing (i.e. dividing by the standard deviation). With a formula, the s th observation of variable Y is transformed into a Z-score as¹³

$$Z_{Y_s} = \frac{Y_s - M_Y}{\sigma_Y} .$$

We need a subscript here because we will use Z-scores for Y and W .

When variables have been transformed into Z-scores, the formula for the correlation becomes conceptually much simpler. However, this is not, in general, a very practical way of computing the coefficient of correlation¹⁴ even if it helps us to understand several of its properties. Specifically, if Z_{Y_s} denotes the Z-score for the s th observation of the variable Y , and Z_{W_s} denotes the Z-score for the s th observation of the variable W , then the coefficient of correlation can be obtained from the Z-scores as

$$r_{W \cdot Y} = \frac{1}{S} \times \sum Z_{W_s} Z_{Y_s} . \quad (2.5)$$

This can be shown, quite simply, by developing Equation 2.5:

$$\begin{aligned} r_{W \cdot Y} &= \frac{1}{S} \times \sum Z_{W_s} Z_{Y_s} = \frac{1}{S} \sum \left(\frac{W_s - M_W}{\sigma_W} \right) \left(\frac{Y_s - M_Y}{\sigma_Y} \right) \\ &= \sum \frac{(W_s - M_W)(Y_s - M_Y)}{S \times \sigma_W \times \sigma_Y} . \end{aligned}$$

$$\text{But, } S \times \sigma_W \times \sigma_Y = \sqrt{\sum (W_s - M_W)^2 \sum (Y_s - M_Y)^2} ,$$

and therefore we get,

$$\begin{aligned} \frac{1}{S} \times \sum Z_{W_s} Z_{Y_s} &= \frac{\sum (W_s - M_W)(Y_s - M_Y)}{\sqrt{\sum (W_s - M_W)^2 \sum (Y_s - M_Y)^2}} \\ &= \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}} \\ &= r_{W \cdot Y} . \end{aligned}$$

¹³ In what follows, we use in the formula for Z-score the population standard deviation σ . We could have used the sample standard deviation ($\hat{\sigma}$) as well. In this case, however, we need to substitute the value $(S - 1)$ for S in the formulas of this section.

¹⁴ So don't use it for computing! Using Z-scores requires more computations and is likely to generate rounding errors.

2.7.1 Computing with Z-scores: an example

To illustrate the computation of the coefficient of correlation using Z-scores, we are going to use again the data from the example 'length of words and number of lines'. The first step is, indeed, to compute the Z-score. For this we need the mean and standard deviation of each variable. From Table 2.2, we find the following values:

- For the means,

$$M_W = \frac{160}{20} = 8.00$$

$$M_Y = \frac{120}{20} = 6.00$$

- For the variances,

$$\sigma_W^2 = \frac{294}{20} = 14.70$$

$$\sigma_Y^2 = \frac{150}{20} = 7.50$$

- For the standard deviations,

$$\sigma_W = \sqrt{\sigma_W^2} = \sqrt{14.70} \approx 3.83$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{7.50} \approx 2.74 .$$

With these values, we can fill in Table 2.3 which gives, in turn, the quantities needed to compute the coefficient of correlation using Z-scores. You can check that

$$\begin{aligned} r_{W \cdot Y} &= \frac{1}{S} \times \sum Z_{W_s} Z_{Y_s} \\ &= \frac{1}{S} \times (-1.7143 + 0 + \dots - 1.3333 - 0.7619) \\ &= \frac{-14.6667}{20} \\ &= -.7333 , \end{aligned}$$

which, indeed, is the same value as was found before.

2.7.2 Z-scores and perfect correlation

Another advantage of expressing the coefficient of correlation in terms of Z-scores is that it makes clear the fact that the values of +1 and -1 correspond to a perfect correlation between two variables. Two variables are perfectly correlated when they vary in the exact same way, or in other words, when the shapes of their distributions are identical. In that case the Z-scores corresponding to the first variable, W , are equal to the Z-scores corresponding to the second variable Y :

$$Z_{W_s} = Z_{Y_s} \quad \text{for all } s .$$

Hence, the formula for r becomes

$$r_{W \cdot Y} = \frac{1}{S} \times \sum Z_{W_s} Z_{W_s} = \frac{1}{S} \times \sum Z_{W_s}^2 . \quad (2.6)$$

Word	Length Y	Number of Lines W	Z_{Y_s}	Z_{W_s}	$Z_{Y_s} \times Z_{W_s}$
bag	3	14	-1.0954	1.5649	-1.7143
across	6	7	0	-0.2608	0
on	2	11	-1.4606	0.7825	-1.1429
insane	6	9	0	0.2608	0
by	2	9	-1.4606	0.2608	-0.3810
monastery	9	4	1.0954	-1.0433	-1.1429
relief	6	8	0	0	0
slope	5	11	-0.3651	0.7825	-0.2857
scoundrel	9	5	1.0954	-0.7825	-0.8571
with	4	8	-0.7303	0	0
neither	7	2	0.3651	-1.5649	-0.5714
pretentious	11	4	1.8257	-1.0433	-1.9048
solid	5	12	-0.3651	1.0433	-0.3810
this	4	9	-0.7303	0.2608	-0.1905
for	3	8	-1.0954	0	0
therefore	9	1	1.0954	-1.8257	-2.0000
generality	10	4	1.4606	-1.0433	-1.5238
arise	5	13	-0.3651	1.3041	-0.4762
blot	4	15	-0.7303	1.8257	-1.3333
infectious	10	6	1.4606	-0.5216	-0.7619
Σ	120	160	0	0	-14.6667

Table 2.3 Raw scores, Z-scores and Z-score cross-products for the example: 'Length of words and number of lines of the definition'. $M_W = 8$, $M_Y = 6$, $\sigma_W \approx 3.83$, and $\sigma_Y \approx 2.74$.

Remember that Z-scores have a standard deviation equal to 1, and therefore a variance also equal to 1 (because $1^2 = 1$). From the formula of the variance of a set of scores, we find that for any set of Z-scores we have the following equality (remember that the mean of the Z-scores is zero):

$$\sigma_Z^2 = 1 = \frac{\sum_s (Z_s - M_Z)^2}{S} = \frac{\sum_s (Z_s - 0)^2}{S} = \frac{\sum_s Z_s^2}{S}.$$

This last equation implies that for any set of Z-scores

$$\sum Z_s^2 = S. \quad (2.7)$$

Now if we plug the result from Equation 2.7 into Equation 2.6, we find that when the shapes of two dependent variables are the same, the value of their coefficient of correlation becomes

$$r_{W.Y} = \frac{1}{S} \times \sum Z_{W_s}^2 = \frac{1}{S} \times S = 1 ,$$

which shows that when the variables Y and W are perfectly correlated, the magnitude coefficient of correlation will be equal to 1.

2.8 Correlation and causality

An important point to keep in mind when dealing with correlation is that *correlation does not imply causality*. The fact that two variables are correlated does not mean that one variable causes the other. There are, in fact, quite a lot of examples showing the consequences of mixing correlation and causality. Several of them can be somewhat silly or funny (it is almost a professional hazard for statisticians to look for these cases). For example, the intellectual development of children as measured by the DQ or developmental quotient, is highly and significantly correlated with the size of their big toes—the higher the intellectual ability, the larger the big toe. Does that mean that intelligence is located in the big toe?

Another example: in France, the number of Catholic churches, as well as the number of schools, in a city is highly correlated with the incidence of cirrhosis of the liver (and the number of alcoholics), the number of teenage pregnancies, and the number of violent deaths. Does that mean that (French) churches and schools are sources of vice? Does that mean that (French) newborns are so prone to violence as to be murderers? Actually, it is more reasonable to realize that the larger a city is, the larger the number of churches, schools and alcoholics, and so on, is going to be. In this example, the correlation between number of churches/schools and alcoholics is called a *spurious* correlation because it reflects only their mutual correlation with a third variable (here the size of the city).

However, the existence of a correlation between two dependent variables can be used as a practical way to predict¹⁵ the values of a dependent variable from another one. For example, Armor (1972, see also Pedhazur, 1982) after a re-analysis of a very important educational survey called the Coleman report (1966), found that a composite index—derived from the possession of several domestic appliances (e.g. TV sets, telephone, refrigerator) and some other items (such as an encyclopedia)—shows a correlation of around .80 with a score of verbal intelligence. As a consequence, this index can be used to predict the verbal score of children (if it is impossible or impractical to measure it directly, of course). But it would be silly to infer from this correlation that a fridge *causes* verbal intelligence (maybe fridges are talkative around children?), or that buying appliances will increase the verbal intelligence of our children (but what about the encyclopedia?).

Correlations between several dependent variables are in general studied using techniques known as factor analysis, principal component analysis, or structural equation modelling (see e.g. Loehlin, 1987; Pedhazur and Pedhazur-Schmelkin, 1991; Jolliffe, 2003). The main goal of these methods is to reveal or express the relations between the dependent variables in terms of hidden (independent) variables often called factors, components, or latent variables.

¹⁵ How to use the coefficient of correlation to predict one variable from the other one is explained in Chapter 4 on regression analysis.

It is worth noting, in passing, how much the naturalistic and the experimental approach differ in the way they consider individual differences. The naturalistic approach essentially relies upon individual differences in order to detect the effects of variables of interest (cf. Chateau, 1972). On the other hand, the experimental approach treats individual differences as a nuisance (i.e. error) and actively tries to eliminate or control them.

However, even though a correlational approach is founded on individual differences, it can be used as a first test (or a *crucible* to use the pretty word of Underwood, 1975) for theories. Actually a good number of theories tested in the laboratory are also amenable to testing with a correlational approach (pre-tested, Underwood would say). For example, if a theory predicts that imagery affects memory, then subjects spontaneously using imagery (or 'highly visual subjects') should exhibit a better level of performance on a memory task than subjects who do not use imagery. As a consequence, we should be able to find a correlation between the dependent variable 'spontaneous use of imagery' and the dependent variable 'memory performance' (both are dependent variables because we are just measuring them, not controlling them). If the correlation between imagery and memory is close to zero, then the theory should be abandoned or amended. If the correlation is high, the theory is not proven (remember the slogan: 'correlation is not causation'). It is at best supported, but we feel more confident in moving to an experimental approach in order to test the causal aspect of the theory. So a correlational approach can be used as a first preliminary step.

2.9 Squared correlation as common variance

When squared, the coefficient of correlation can be interpreted as the *proportion of common variance* between two variables. The reasoning behind this interpretation will be made clear after we have mastered regression analysis, so at this point we will just mention this property. For example, the correlation between the variables length (W) and number of meanings (Y) is equal to $r_{W.Y} = -.7333$ (cf. Equation 2.3), therefore the proportion of common variance between W and Y is equal to $r_{W.Y}^2 = (-0.7333)^2 = .5378$. Equivalently, we could say that W and Y have 54% of their variance in common. Because a coefficient of correlation takes values between -1 and $+1$, the squared coefficient of correlation will always take values between 0 and 1 and its magnitude will always be smaller than the magnitude of the coefficient of correlation.

Chapter summary

2.10 Key notions of the chapter

Below are the main notions introduced in this chapter. If you have problems understanding them, you may want to re-read the part(s) of the chapter in which they are defined and used. One of the best ways is to write down a definition of each of those notions by yourself with the book closed.

Linear relationship
Scatterplot

Cross-product
Covariance

Pearson coefficient of correlation
Perfect correlation

Correlation and causality

2.11 Key formulas of the chapter

Below are the main formulas introduced in this chapter: try to go through them and understand what they mean.

$$SCP_{WY} = \sum_s (W_s - M_W)(Y_s - M_Y)$$

$$COV_{WY} = \frac{SCP}{S}$$

$$r_{W \cdot Y} = \frac{COV_{WY}}{\sigma_W \sigma_Y}$$

$$= \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}}$$

$$= \frac{1}{S} \times \sum Z_{W_s} Z_{Y_s}$$

2.12 Key questions of the chapter

Below are some questions about the content of this chapter. All the answers are to be found in the chapter. If you are in any doubt about your answer, you may want to re-read parts of the chapter.

- * What types of information can you derive from a scatterplot ?
- * What is the goal of the coefficient of correlation?
- * In which case would you use a correlation analysis?
- * What can we conclude from a positive value of the coefficient of correlation and why?
- * What can we conclude from a negative value of the coefficient of correlation and why?
- * What can we conclude from a zero value of the coefficient of correlation and why?
- * Why do we use the coefficient of correlation rather than the cross-product or the covariance to assess the relationship between two variables?
- * What are the limitations of the coefficient of correlation?
- * There is a strong correlation between the number of fire trucks sent to fires and the amount of damage done by those fires. Therefore, to minimize fire damage we should send fewer fire trucks to fires. Are you convinced?