JOINT DISTRIBUTIONS

Discussion until now has focused on distributions of single variables looked at on their own - univariate descriptive statistics. However, it is often interesting in psychology to ask if the distributions of **two** variables are related to each other; for instance, if the distribution of VCE marks is related to **IQ** scores. Is there, for instance, a tendency for people who have high IQs to get high VCE scores, those who have low IQs to get low scores?

We are now dealing with bivariate descriptive statistics. The question here is whether or not IQ scores and VCE scores covary or, to put it slightly differently, whether the criterion of VCE score can be predicted from the predictor IQ. Table 8.1 shows the IQs and VCE scores of 20 Bendigo students. We can now look at these scores to see whether they covary, and to what extent they covary.

SCATTER DIAGRAMS

The simplest way to look at this issue is to construct a scatter diagram, a simple graph. We start by identifying one of the variables as the predictor, the other as the criterion. The predictor may also be referred to as the independent variable, the criterion as the dependent variable.

We then graph the predictor scores against the criterion scores, putting the predictor values along the *x* -axis, the criterion scores along the y-axis. Each person becomes a point on the graph, the pair of predictor and criterion scores for that person defining his or her coordinates on the graph. The IQs in Table 8.1 could be graphed along the x- axis, VCE scores along the y-axis.

Table 8.1: IQs and VCE scores of Bendigo students

PERSON	X_t (IQ)	Y_i (VCE Score)
1	117	298
2	132	348
3	109	270
4	121	285
5	130	290
6	107	272
7	124	310
8	127	315
9	118	290
10	115	288
11	140	360
12	120	310
13	111	273
14	132	345
15	124	320
16	114	278
17	117	290
18	120	295
19	107	275
20	115	288

For this reason, it is common to refer to the predictor (the independent variable) as X and the criterion (the dependent variable) as Y. The coordinates of person 1 would then be 117 and 298 (see Table 8.1), and the person would be placed on the graph as a dot having the -coordinate 117 and the y-coordinate 298 (see Figure 8.1).

Figure 8.1: Person 1 from Table 8.1 placed as a point on a graph

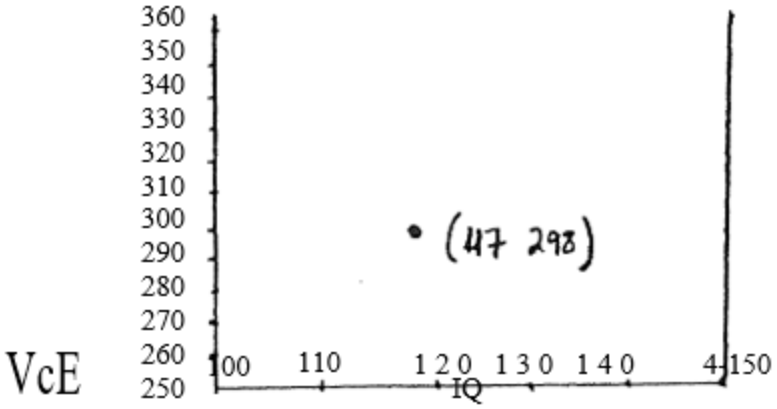
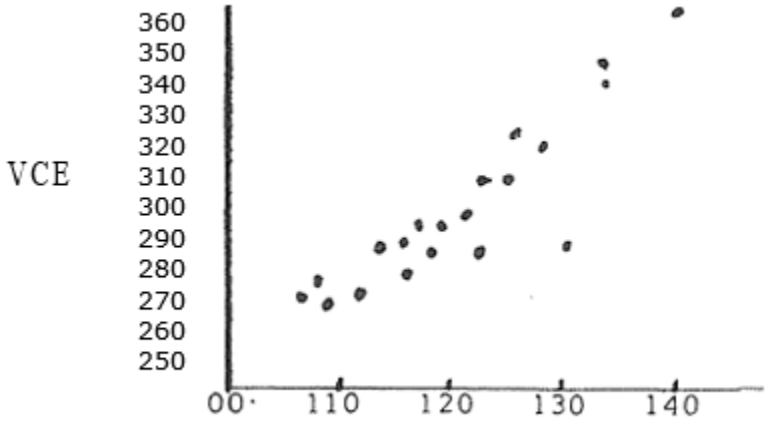


Figure 8.2: Scatter diagram of Bendigo students for IQ and VCE scores



As can be seen, the bivariate distribution yielded by these scores defines a cloud or swarm of points, which in this case in long and narrow and at an angle of about 45 degrees to the x-axis. A cloud of this shape means that there is a close relationship between IQ and VCE scores, i.e., it is possible to predict the VCE scores of this group of students from their IQs with a high degree of accuracy, because they covary closely (their covariance is high - variance you already know about from Section 6.4, "co" is from the Latin for "with" or "together" - IQs and VCE scores vary together). Furthermore, the relationship is said to be positive, because high IQs go with high VCEs, low IQs with low VCEs, intermediate scores with intermediate VCEs, and so on. Figure

8.3 shows a high negative relationship, while Figure 8.4 shows the scatter diagram for two variables which have no relationship, to each other (i.e., it is not possible to predict the scores on one variable from those on the other).

Figure 8.3: Scatter diagram for two variables which covary strongly and negatively

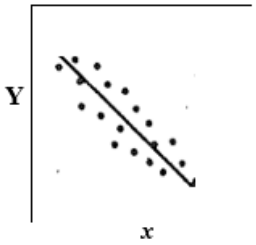
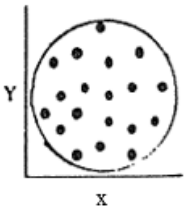


Figure 8.4: Scatter diagram for two variables which do not covary

Figure 8.4: Scatter diagram for two variables which do not covary



The thinner the ellipse defined by the scatter diagram, the stronger the degree of covariation between the two variables. The fatter the ellipse, the weaker the covariation. When the scatter diagram defines a circle, there is no covariation.

Linear and nonlinear relationships

There is no guarantee that the points in a scatter diagram will define an ellipse, as in the examples given in Figure 8.2 and 8.3. On those occasions when they do this, the relationship between the two variables is said to be **linear.** Thus Figures 8.2 indicates that there is a linear relationship between IQ and VCE score in the sample of Bendigo students. The relationship in Figure 8.3 is also linear. A linear relationship means that increasing IQ scores are related to increasing VCE scores **at all 10 levels.** Suppose that increased IQ scores related to VCE scores only up to an IQ value of 125, and that thereafter increases in IQ yielded no change in VCE. This form of covariation would yield the scatter diagram shown in Figure 8.5. Such a scatter diagram represents a relationship which is said to be **curvilinear.** Figure 8.6 shows other forms of curvilinear relationships.

Figure 8.5: A curvilinear scatter diagram

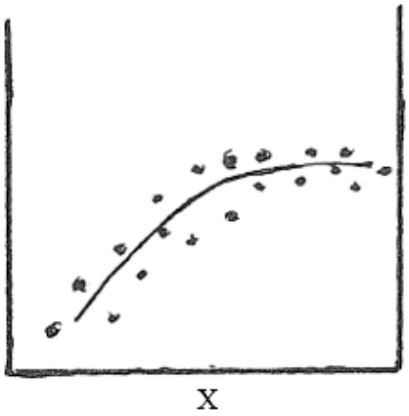
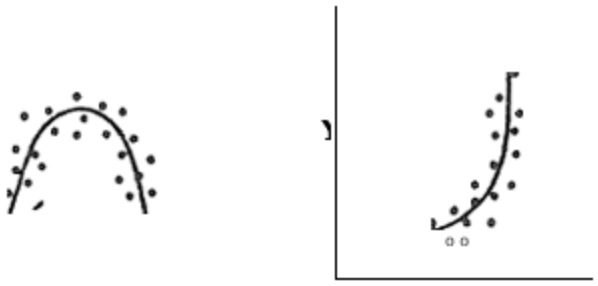


Figure 8.6 Other curvilinear relationships



Covariation and Correlation

The tendency for scores on two variables to vary with each other is referred to as **covariation.** When two variables covary this means that if you know how a group of subjects distribute themselves around the mean on one variable (i.e., if you know the variance of the one variable), you automatically know something about the way **the same subjects** are distributed around the mean on the second variable. This is the state of affairs when the scatter diagram is any shape other than a circle--to simplify matters, we will stick to linear relationships here (see Figure 8.2 and 8.3). If the variables X and Y covary, you can, for instance, make predictions about scores on variable Y on the basis of a knowledge of the X scores, and do better than simply guessing.

To stick to the example of IQs and VCE scores, covariation of these scores would mean that a knowledge of IQs would make it possible to make predictions about VCE scores; the larger the **covariance** the more accurate the predictions.

Correlation

In the case of linear relationships between variables, the tendency for scores to go together is referred to as correlation. When the covariance is high, the variables correlate strongly or highly. If high scores on the predictor go with high scores on the criterion, low with low, the correlation is positive, if high scores on the one variable go with low scores on the other, the correlation is negative. The degree of correlation is expressed in the form of a correlation coefficient; this coefficient ranges from 0.00 (no correlation at all) to 1.00 (perfect correlation--you can predict criterion scores perfectly from the predictor). If the correlation is perfect but in the reverse direction (i.e., high scores on the one variable go with low scores on the other), the correlation coefficient is -1.00 (The + or - sign in front of the correlation coefficient tells you nothing about the strength of the correlation, but only the direction: a + sign means that high scores go with high scores, low with low; a - sign means that high scores go with low scores, low with high).

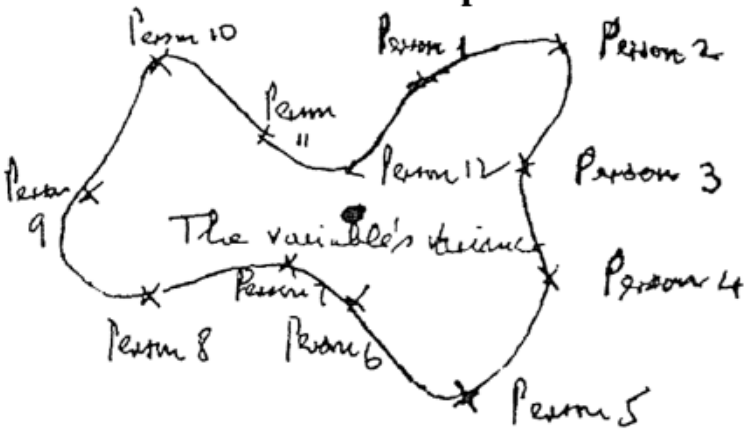
Where the strength of the relationship between the two variables is greater than 0.00 but less than 1.00 (note that in the case of positive correlations it is usual to leave out the + sign), the correlation coefficient will have a value greater than 0.00 but less than 1.00: a correlation coefficient of, let us say, 0.75 means that there is a decided tendency for scores to go together, but that it is less than perfect; a correlation coefficient of 0.25 indicates a stronger than zero relationship, but one which is quite weak.

To return to the example of the possible relationship between IQs and VCE scores, a correlation of +1.00 would mean that it is possible to make perfect predictions of VCE marks from IQs, the person with the highest IQ getting the highest VCE score, the second highest IQ the second highest VCE score, and so on. A coefficient of -1.00 would also mean perfect prediction, except that the person with the highest IQ would get the lowest VCE score, the second highest IQ the second lowest VCE score, and so on. The degree of correlation between IQs and VCE scores shown in Figure 8.2 is positive (high goes with high) and substantial (because the swarm is long and thin). It represents a correlation coefficient of about +0.83.

The coefficient of determination

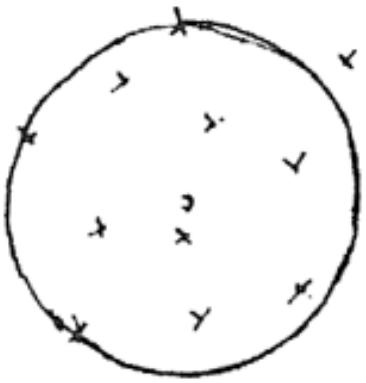
A useful way of conceptualising covariation, and hence correlation, since correlation results from covariation, is to think of a distribution as involving a cloud of points about a central point. The central point is the mean, and the distances of the individual scores are the distances of the points from the mean. The area of the cloud represents the variance of the variable. It is important to note that this way of representing variance is simply a metaphor—a way of making the concept easily understandable not an exact mathematical representation. Figure 8.7 shows a distribution represented in this way.

Figure 8.7 The distribution of a variable represented as a cloud



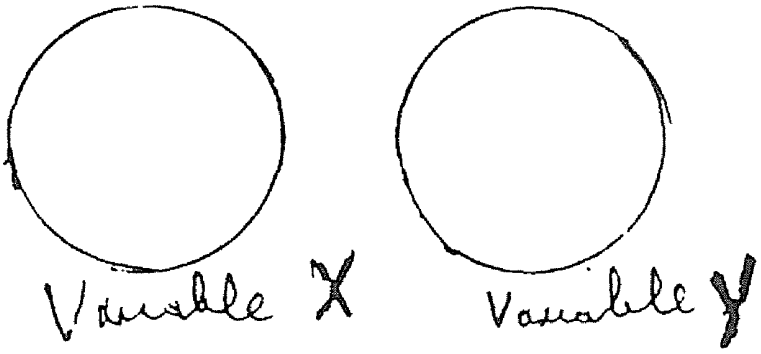
If the individual scores are arranged in such a way as to produce the most symmetrical cloud and the curve is smoothed out, the variance can be represented as a circle, as in Figure 8.8.

Figure 8.8: Variance depicted as a circle



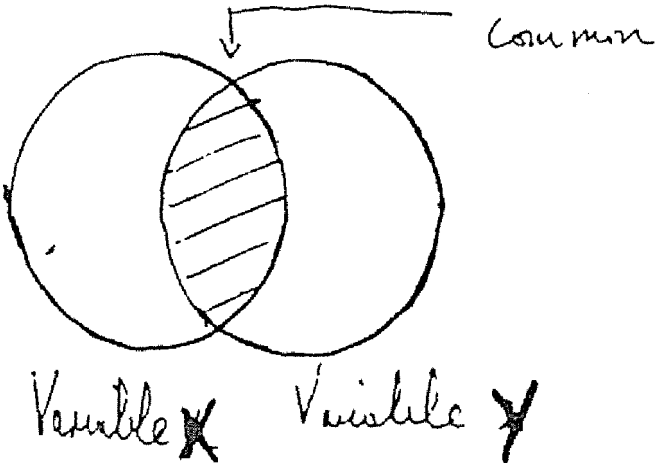
If the scores on the variable in question are transformed into z scores, this circle always has an area of 1.00 (remember that z scores always have a standard deviation and a variance of 1.00). Consider now any pair of variables (on each variable in the pair we have the scores of the **same subjects);** in z score form their variances could be depicted as two circles, each with an area of 1.00. Where the two variables have a correlation of 0.00 (they are uncorrelated), the situation would be as shown in Figure 8.9.

Figure 8.9: Two uncorrelated variables



However, in the case of correlated variables, the position is as shown in Table 8.10.

Figure 8.10: Geometric representation of correlated variables



The higher the correlation between the two variables, the greater the proportion of common variance (the higher the covariation). In the case of perfect correlation, the two circles would overlap exactly. The shaded area shown in Figure 8.10 becomes larger as the degree of correlation increases, smaller as the degree of correlation decreases. In fact, the correlation coefficient is the square root of the covariance. If the shaded area were 0.4, as is approximately the case in Figure 8.10 (it is about 40% of the area of either circle), the correlation coefficient would be 1 0.4, and this is 0.63, i.e., Figure 8.10 is a graphical representation of the situation where the two variables in a bivariate distribution correlate 0.63 (this form of representation says nothing about the sign of the correlation coefficient).

More commonly, the correlation coefficient between two variables is already known, and the unknown factor is the degree of covariance. In this case, the formula is the other way round: the proportion of common variance is equal to the correlation coefficient squared. This statistic, the proportion of common variance, is referred to as the coefficient of determination (the extent to which the variance of one variable is determined by the other one) and is calculated by squaring the correlation coefficient. It is customary to use the symbol "r" to represent the correlation coefficient.

Correlation coefficients for different kinds of data

You may recall the discussion of different kinds of data (see Section 2 starting on page 2) - nominal data, ordinal data, interval data, ratio data. For each of these kinds of data there are special procedures for calculating correlation coefficients. The most common of these are set out below.

1. Interval and Ratio Data

The "prototypical" correlation coefficient is the Pearson Product-Moment Correlation Coefficient. The coefficient is used when calculating correlations between two variables, both of which involve interval or ratio data (for this purpose no distinction is made between interval and ratio). The example already given (IQ coefficients and VCE scores) involves data with which it is appropriate to calculate Product-Moment coefficients. The Product-Moment correlation coefficient is represented by the symbol r: other kinds of correlation coefficients have their own symbols. (The explanation of covariance and coefficient of determination already given holds true for the Product-Moment coefficient and also for two other correlation coefficients which will be outlined in later paragraphs.)

The formula for calculating the Product-Moment correlation coefficient is based on the covariance of two variables. The covariance is divided by the product of the Standard Deviations.

Ordinal Data

It is not uncommon in psychological research to obtain data which do not meet the requirements for interval data. There is a special formula for calculating the degree of covariance between variables where the data are in the form of ranks (i.e. ordinal data). The resulting correlation coefficient is called Spearman's, rank-order coefficient and is abbreviated p (the Greek letter r

Nominal Data

Suppose we were interested in the extent to which being a smoker or a non-smoker is related to gender: smoker v. non-smoker is a nominal data (you are classified either the one or the other) and gender is similar. With such data it is possible to calculate a number of coefficients indicating the degree to which variables go together. The exact coefficient used depends on whether both variables are truly **dichotomous** (can have only two values) or result from collapsing multicategory data into dichotomies (for instance, dividing human beings into two "artificial" categories "old" versus "young"). It also depends upon whether the nominal data are dichotomies or have more than two categories (for instance "nonsmoker", "former smoker", "light smoker", "heavy smoker", or "young", "middle aged", "old").

Fisher's z transformation of correlation coefficients

As has already been pointed out, the proportion of covariance shared by two correlated variables is determined by **squaring** the correlation coefficient (coefficient of determination). This means that the relationship between correlation coefficients and proportion of common variance is **nonlinear**. Equal changes in the correlation coefficient do not mean equal changes in the amount of variance accounted for. (Remember that for interval and ratio data equal differences in the numbers on the scale must indicate equal differences in whatever is being measured.) For instance, the difference in height between 1.64m and 1.69m (5cm) is exactly the same as the difference between 1,78m and 1.83m (also 5cm). Because of the fact that the amount of variance accounted for is equal to the square of the correlation coefficient, however a change from a correlation of, let us say, 0.35 to 0.48 (a change of 0.13) does not mean the same increase in the variance accounted for as a change from 0.64 to 0.77, despite the fact that the change in the correlation coefficients is the same. In fact, an increase in the correlation coefficient from 0.35 to 0.48 involves an increase (of 88.1%) in variance accounted for (an increase from 12.25% to 23.04%), whereas an increase from 0.64 to 0.77 results in an increase in variance accounted for of 44.8% (from 40.96% to 59.29%). These correlation coefficients do not constitute interval data.

In order to be able to calculate means and similar statistics with correlation coefficients, it is necessary to **transform** them into data complying with the requirements for interval data. The formula for the transformation, which I will not give here, was worked out by Fisher, and the transformed correlation is referred to as **Fisher's z.** The abbreviation for Fisher's z is written Z_r (remember that r is the customary abbreviation for correlation coefficients). Many statistics textbooks have a table for transforming r into z_r, and it is usual to look up the values rather than calculating them oneself. Transformation to makes it possible to work out the mean of a group of correlation coefficients or their standard deviation.