Chapter 2

Bivariate analyses

Although it has been assumed in preparing this book that readers have some background in the simpler statistical methods and their use in geographical research, two of these methods are reviewed in the present chapter. Bivariate correlation and regression and the one-way analysis of variance are the simplest cases of the family of procedures which are covered in the rest of the book so treatment of them allows development of a series of building-blocks which are used throughout the remaining chapters.

Correlation and regression

The geographer's question 'Are there relationships between phenomena in various locations?' is typical of the more general statistical question enquiring into relationships among *variables* (characteristics) over a set of *observations* (places). The method used to answer it depends on the form of the available data. If they are on an interval- or a ratio-scale (p. 8), then correlation and regression analysis may be used, as long as certain other conditions are met (p. 37).

A hypothetical problem

In formal terms, we state a hypothesised relationship that

$$Y = f(X) \qquad (2.1)$$

which is a shorthand way of arguing that the value of Y at a particular observation is some function of the value of X there. (In other words, Y might be 2X, or twice the value of X; it might be 0·5X, 2·76 $\log_{10}$X, and so on.) If this were a perfect function, then from knowledge of X at an observation, we could estimate the value of Y with certainty. Such functions are illustrated in Fig. 2.1.

Because the functional relationships portrayed in Fig. 2.1 are perfect, all individual observations fall on the lines representing those functions. (Only straight-line functions are shown in Fig. 2.1; non-linear relationships are also feasible, and are dealt with later – p. 38.) The laws of nature, including those of human nature, whose effects and operations we study, very rarely, if ever, produce a completely predictable result, however. In some cases, this is because of imperfections in our measurement

Table 2.1 Educational needs and spending: hypothetical data

Town	X*	Y†	Town	X	Y	Town	X	Y
1	17	27	17	47	57	34	27	31
2	42	64	18	62	70	35	52	66
3	49	52	19	45	50	36	59	69
4	35	59	20	36	54	37	33	42
5	69	77	21	48	45	38	27	43
6	48	54	22	50	68	39	33	60
7	34	50	23	30	45	40	31	46
8	43	48	24	44	58	41	49	46
9	39	49	25	24	36	42	42	52
10	56	72	26	39	40	43	40	46
11	38	52	27	64	60	44	49	58
12	43	62	28	50	64	45	42	58
13	44	50	29	41	51	46	29	40
14	57	50	30	45	36	47	43	55
15	36	40	31	37	47	48	53	63
16	22	34	32	51	38	49	39	59
			33	54	44	50	33	56

*X = percentage of the population aged between 5 and 21

†Y = *per capita* expenditure on education

procedures, but in most it is because effects are usually the product of a variety of causes operating in conjunction, so that a single independent variable will not allow us to estimate the value of Y with complete certainty. And so instead we have to study general trends in our data sets, providing estimates of the relationship between X and Y implied in formula (2.1) and also of the accuracy with which the values of Y can be derived from this estimated relationship.

If we square the two components of $(Y_i - \overline{Y}_T)$ in formula (2.51) we get

$$(Y_i - Y_T)^2 = (Y_i - \overline{Y}_G)^2 + 2(Y_i - \overline{Y}_G)(\overline{Y}_G - \overline{Y}_T) + (\overline{Y}_G - \overline{Y}_T)^2 \quad (2.52)$$

and summing this for all towns produces the *total variation*,

$$\sum_{i=1}^{N} (Y_i - \overline{Y}_T)^2 = \sum_{i=1}^{N} (Y_i - \overline{Y}_G)^2 + \sum_{i=1}^{N} 2(Y_i - \overline{Y}_G)(\overline{Y}_G - \overline{Y}_T) + \sum_{i=1}^{N} (\overline{Y}_G - \overline{Y}_T)^2 \quad (2.53)$$

Because $\Sigma (Y_i - \overline{Y}_G) = 0$ (i.e. the positive and negative deviations around

To illustrate how we set about this task, a hypothetical data set has been constructed to test a hypothesis that the greater the demand for education in a place, the more the money that will be spent providing for it. To investigate the validity of this proposition, demand is measured as the percentage of the towns population aged between 5 and 21 (this is the independent variable, X) and expenditure is the amount spent per resident of the town on the education budget (the dependent variable, Y). Our expectation is that *as the percentage of the population aged between 5 and 21 increases so the amount spent per resident increases*.

Applying these formulae to the data of Table 2.2 produces

$$\Sigma (Y_{iF} - \overline{Y}_F)^2 = 945{\cdot}46$$

$$\Sigma (Y_{iS} - \overline{Y}_S)^2 = 1036{\cdot}10$$

$$\Sigma (Y_{iL} - \overline{Y}_L)^2 = 1219{\cdot}44$$

(in all these cases summation is over the relevant group memberships only), so that

$$\sum_{i=1}^{N} (Y_i - \overline{Y}_G)^2 = 3201{\cdot}00$$

which is the with in-groups variation.

$$\sum_{i=1}^{N} (Y_i - \overline{Y}_T)^2 = 6038{\cdot}02$$

The ratio in formula (2.55) can be modified so that it is comparable to the product moment correlation coefficient, r_{YX}. In a regression analysis, total variation is $\Sigma (Y_i - \overline{Y})^2$ and residual variation, unaccounted for by the regression, is $\Sigma (Y_i - \hat{Y})^2$, according to formula (2.18). The proportion of the variation unaccounted for by the independent variable is thus

$$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{N} (\overline{Y}_i - \overline{Y})^2 = 1 - r_{YX}^2 \quad (2.56)$$

In analysis of variance, the independent variable is group membership so that

$$\sum_{i=1}^{N} (Y_G - \overline{Y}_T)^2 \Big/ \sum_{i=1}^{N} (\overline{Y}_i - \overline{Y}_T)^2 \quad (2.57)$$

is the proportion of the variation in Y accounted for by the classification of observations into types, whereas the ratio of the within groups to total variation

$$\sum_{i=1}^{N} (Y_i - \overline{Y}_G)^2 \Big/ \sum_{i=1}^{N} (\overline{Y}_i - \overline{Y}_T)^2 \quad (2.58)$$

is the residual or unaccounted for variation. The equivalent of $(1 - 1\ r_{YX}$ in correlation analysis is thus given by formula (2.58), whereas formula (2.57) presents an equivalent to the squared correlation coefficient, r_{YX}.

Data for these two variables are available for 50 towns (Table 2.1). An immediate indication of the relevance of the hypothesis is obtained by plotting the 50 pairs of values on a scattergram (Fig. 2.2) on which, by convention, X forms the horizontal and Y the vertical axis. The general trend is as expected. For a precise statement, however, we need to know the functional relationship, or how well the regression line fits the pattern of dots on the scattergram: this is given by a correlation coefficient. The amount of change in the pattern