3

Key concepts 2
Statistics

3.1 Introduction

This chapter is concerned with ways of summarizing and exploring numerical data. Even a brief summary of the key principles of statistics would require a dedicated book, so the intention of this chapter is to introduce some (very selective) ideas that it is necessary to understand to make use of parts of the rest of this book. These methods provide the basis of the *spatial* statistical methods that will be defined later on. The analysis of aspatial data (data with no spatial locational information) and spatial data usually starts with computation of standard summary statistics, as described in this chapter. Statistics can be divided into descriptive statistics, which provide summaries, and inferential statistics, which allow the making of inferences about a population (a complete data set representing all cases, e.g. all people in a country) from a sample. A sample is a partial data set, such as a population data set which excludes some people for some reason such as cost limitations, enabling only a limited survey. Both descriptive statistics and inferential statistics are introduced in this chapter, although more space is devoted to the former. Core concepts, which will be discussed in Section 3.4, include probabilities and the significance level. A statistical hypothesis may be associated with a probability that it is true or false and this is a central notion in statistics.

The following sections consider the purpose of statistical methods and introduce some ways of describing data sets. The focus here is initially on univariate statistics—methods that are used to analyse only one variable. Next, the focus is on multivariate methods—methods that deal with two or more variables simultaneously. In addition to introducing methods, the chapter will introduce some of the principles of statistical notation, for example one concern is to demonstrate how to 'read' the equations given

in the rest of the book. Material of this nature is sometimes placed, for good reason, in the appendices of introductory books. In this book, such material is presented within the main text so that there is a direct transition from standard aspatial statistics (i.e. methods that do not take into account the spatial location of observations) to their spatial equivalents. This chapter deals exclusively with standard aspatial statistical methods. Methods which do take spatial location into account are the subject of later sections and Section 4.8 deals with the principles of one statistical approach to characterizing spatial variation (i.e. geographical patterning) in the property of interest.

Before proceeding, it is useful to consider how the kinds of data we have to work with may differ. Data may be divided into four types, which contain different amounts of information. These data types are:

Nominal An arbitrary naming scheme, for example ethnic group (White, Caribbean, African).

Ordinal Values are ordered, but there is no information on the relative magnitude of values, for example small, medium, large.

Interval The intervals between measurements are meaningful, but there is no natural zero point, for example temperature. Differences between adjacent values are equal, i.e. 28–27 is the same as 99–98. Temperature (where zero is arbitrary, e.g. the freezing point of water for degrees Celsius) is often cited as an example of an interval variable (see, for example, Ebdon, 1985; O'Sullivan and Unwin, 2002). For two temperatures in degrees Celsius (e.g. 10°C and 20°C) and degrees Fahrenheit (50°F and 68°F), the ratios between the two sets of values are different: $^{10°C}/_{20°C} = 0.5$ and $^{50°F}/_{68°F} = 0.74$.

Ratio Values with a natural zero point, ratios as well as intervals, are meaningful. For example, the ratio of 25 to 50 mm is the same as the ratio of the same measurements in inches (i.e. $^{25\,cm}/_{50\,cm} = 0.5$ and $^{9.8\,in}/_{19.7\,in} = 0.5$).

The main concern in this chapter is with the analysis of interval and ratio data and this is also the main focus in Chapters 8, 9, and 10.

3.2 Univariate statistics

The principal focus in this section is on what are termed 'descriptive statistics'—that is, methods to summarize or describe observations (measurements of some property). Summarizing an individual variable (e.g. precipitation amount) is done with reference to its distribution. The distribution of a variable refers to the set of values ordered from the smallest to the largest. Often, identical or similar values are grouped together, for example values 0–10 may be grouped, then values 11–20 and so on. In this way, we can refer to the frequency of values. For example, are most values very small with only a few large values or is there an even proportion of small and large

values with most values being somewhere in between? When values are grouped into classes they can be depicted using a histogram. This is a form of chart that has bars of a size that is in proportion to the number of values in a given class. For example, if there are five observations in class 1 and 10 observations in class 2 then the bar representing class 2 will be twice as high as the bar representing class 1. Figure 3.1 shows a histogram with the range (minimum and maximum) of values represented by each bar indicated (e.g. values in the first bar range from 27.1 to 29). Frequency indicates the number of cases in a bin or class. As an example, there are 10 cases in the range 37.1 to 39.

Using too few or too many classes will not enable representation of important features of the distribution. The number of classes is likely to be determined as a function of the number of observations and the range of values that they take. If there are many thousands of observations (as might be the case, for example, using a remotely sensed image), and there is a sufficiently wide range of values, then it may be sensible to have a large number of classes, producing a 'smoother' distribution than would be possible for a smaller number of observations.

The most common way of summarizing a data set is to compute some kind of average, the mean average is the most well known. Averages are measures of central tendency in a distribution—in some sense the 'middle' value in the distribution. The mean average of the variable, and the notation used to represent this, is detailed below. First, a value of the variable is indicated by z_i. The value itself is given by z, and i is an index—it indicates the observation number. For example, if there are five observations

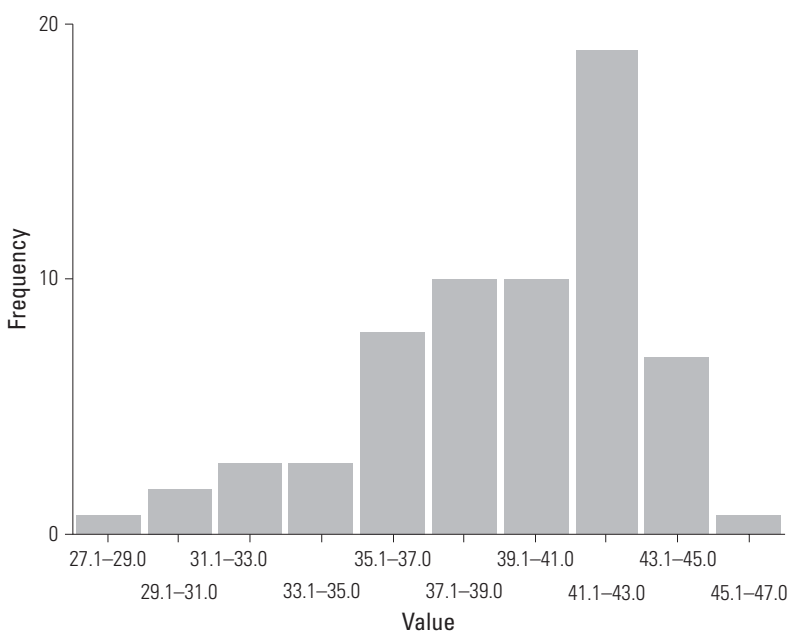

Figure 3.1 Histogram. Values are precipitation amounts in millimetres.

in the data set then i could take a value of 1, 2, 3, 4, or 5. The number of observations is indicated by n and, in the example given, $n=5$. The population (where population indicates it is assumed we have all possible values and not just a sample) mean average, μ (the Greek lower case letter mu), can be given by:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{3.1}$$

The only term not yet explained is Σ (the Greek upper case letter sigma). Σ indicates summation. Below Σ is the term $i=1$ and above it is n. This means start at the first observation ($i=1$) and step through all values up to and including the last value ($i=n$). Note that other sources may use other letters for the variables, but the use of letters and symbols here is consistent throughout the text. $\sum_{i=1}^{n} z_i$ indicates that all values of z_i should be added together.

In the example above z_1 is taken first, then z_2 is added to it and so on until all values have been added together. The end result is then multiplied by $1/n$ (1 being the numerator (top part of the fraction) and n the denominator (bottom part of the fraction)). This gives the mean average of the values and is the same as dividing the summed values by n. As an example, if we have five values (z_1 to z_5) and they are 11, 14, 13, 9, and 6 then their sum is 53, $1/5=0.2$ and the mean average is given by $0.2\times53=10.6$.

Given these values, Equation 3.1 can be given as:

$$\mu = \frac{1}{5}\sum_{i=1}^{5} z_i = 0.2\times(z_1+z_2+z_3+z_4+z_5) = 0.2\times(11+14+13+9+6) = 0.2\times53 = 10.6$$

Other averages include the median (the middle value when all values are ordered from smallest to largest) and the mode (the most frequently occurring value). When there is an even number of values, the median is the mean average of the two values in the middle of the distribution (e.g. values 40 and 41 out of a total of 80 values with values ordered from smallest to largest). The mean average is very sensitive to outliers (i.e. unusually large or small values) and one benefit of using the median or mode is that the impact of outliers is reduced or non-existent.

The dispersion of a distribution is often of interest, i.e. how much variation is there in the values? The range—the absolute difference between the minimum and maximum value—is one simple measure. As noted above, the median is the middle of the ranked values. The value that falls 25% of the way along the list of ranked values (e.g. the mean average of values 20 and 21 of a total 80 values) is called the lower quartile and the value that falls 75% of the way along the ranked list (e.g. the mean average values of 60 and 61 of a total 80 values) is called the upper quartile. Together, the minimum, lower quartile, median, upper quartile, and maximum provide a summary of the distribution.

The dispersion around the mean—the degree to which values are close to the mean average—is given by the standard deviation. The standard deviation is small when the

values are all quite similar to the mean and large when some values deviate markedly from the mean. The population standard deviation, indicated by σ (lower case sigma), is given by:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \mu)^2} \tag{3.2}$$

Most of the notation is familiar from the equation for the mean average. In this case, the mean (μ) is subtracted from each value and the product (i.e. the outcome) of this subtraction is squared. The squared differences are added together. Once this is done the sum is multiplied by $1/n$ and the square root is taken of this value. In words, the standard deviation is the square root of the average squared difference between observed values and their mean average. The squaring is necessary because if the difference between each value and its mean is not squared, then the sum of differences will be zero. Where the square root is not taken the resulting value is called the variance.

The mean and standard deviation as defined above are population statistics. In recognition of the fact that usually we have only a sample, an alternative form of the standard deviation is computed. The sample mean, $\bar{z}$ (z with a bar on top), is computed as above (i.e. the population mean in Equation 3.1). The sample standard deviation, s, is computed in the same way as in Equation 3.2 except that $1/n$ is replaced by $1/(n-1)$. The reason for this requires some explanation. The mean must be computed before we can compute the variance and a quantity known as the number of degrees of freedom is n minus the number of parameters (such as the mean) estimated, thus $n-1$ in this case (see O'Sullivan and Unwin (2002) for a further account). Another way of putting this is that one degree of freedom is used up in estimating the mean and if we know the mean then we only need $n-1$ values to calculate the value of the nth sample and thus know all values (i.e. if we have five values in total then we need only four values and the mean to work out the fifth value) (Rogerson, 2006). The sample standard deviation is given by:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{3.3}$$

Note that population statistics are by convention given by Greek characters (e.g. σ) and sample statistics by ordinary lower case letters (e.g. s).

Using the same data as before (11, 14, 13, 9, and 6 with a mean average of 10.6), the sample standard deviation is calculated using:

$$\begin{aligned}\sum_{i=1}^{5}(z_i - \bar{z})^2 &= (11-10.6)^2 + (14-10.6)^2 + (13-10.6)^2 + (9-10.6)^2 + (6-10.6)^2 \\ &= 0.16 + 11.56 + 5.76 + 2.56 + 21.16 = 41.20\end{aligned}$$

Given this, s is obtained:

$$s = \sqrt{\frac{1}{5-1} \times 41.20} = \sqrt{0.25 \times 41.20} = 3.21$$

The mean and standard deviation are useful measures of a distribution if it is normal. A normal distribution is characterized by an equal proportion of small and large values, with a peak of values in the middle ranges—the distribution is symmetric. This type of distribution is called bell-shaped. A distribution with a large number of small values and a small number of large values is termed 'positively skewed', while a distribution with a small number of small values and a large number of large values is termed 'negatively skewed' (an example is the histogram in Figure 3.1). The mean average is 'pulled' in the direction of the skew, i.e. it is affected by extreme values. In a skewed distribution, the mode will be under the peak of values, the mean will be closer to the 'tail' of extreme values, and the median will be in between the mode and the mean. The degree of skewness can be measured by a statistic called the coefficient of skewness (note that different measures of skewness exist; the measure below is as implemented in Microsoft® Excel®). This can be given by:

$$\text{skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{z_i - \overline{z}}{s} \right)^3 \tag{3.4}$$

In words, the right-hand side of Equation 3.4 is the sum of the cubed product of differences between individual values and their mean average divided by the standard deviation. Positive values of the skewness coefficient indicate positive skew and negative values indicate negative skew, while a value of zero indicates no skew. Some examples are given in Figure 3.2.

Some distributions have two or more modes, i.e. peaks of values. It is important to examine the distribution of a variable prior to further analysis. Examples of distributions which are normal, positively skewed, and negatively skewed are given in Figure 3.2; for the purposes of the discussion the data are treated as measurements of precipitation amount in millimetres. If a distribution is normal (and we can perceive it as a bell-shaped smooth curve, rather than a histogram with discrete bars, as shown in Figure 3.1) then 68.26% of the values in the data set should fall within one standard deviation of the mean. In other words, 68.26% of the area under the normal curve lies within one standard deviation of the mean (i.e. above or below the mean), 95.46% of the area lies within two standard deviations, and 99.73% lies within three standard deviations. If the distribution is normal, the mean is 10.6, and the standard deviation is 3.21 then 68.26% of the values should be 10.6 ± 3.21 (i.e. in the range 7.39 to 13.81).

Following the discussion above, the mean and standard deviation are not representative if the distribution is not close to normal—this potentially affects many of the procedures detailed in Chapters 8, 9, and 10 in particular. Various possible solutions exist: the variables can be transformed (e.g. by taking the square root or the log of the values) and the transformed variables may have a less skewed distribution. Analysis can

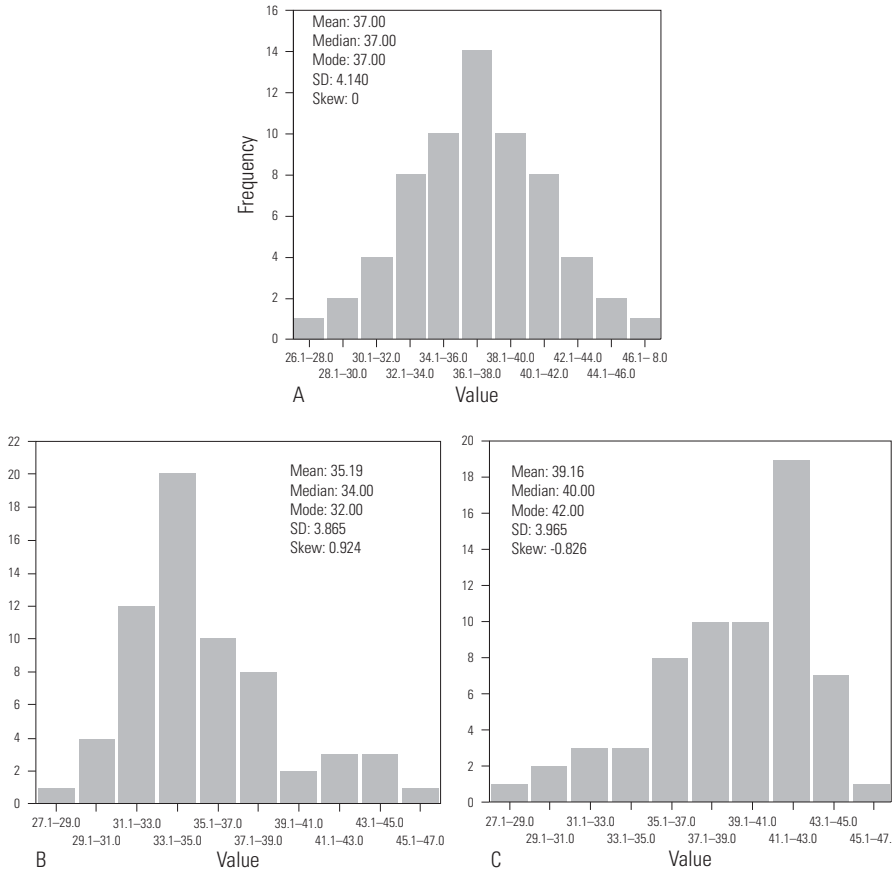

Figure 3.2 Histograms: (A) normal distribution, (B) positive skew, (C) negative skew. Values are precipitation amounts in millimetres. SD is the standard deviation and Skew is the skewness coefficient.

then be conducted using the transformed variables and these can be back-transformed (i.e. converted back to the original values) after the analysis. The logarithmic transformation is used widely to account for distributions with long tails of positive values (i.e. a small proportion of large values); in such cases, the distribution of the log-transformed data should be approximately normal and the transformed data can then be analysed in the usual way. A logarithm of a number is that particular number expressed as a power of another number. Natural logarithms are numbers expressed as powers of e (often given by exp), which is the exponential constant (approximately equal to (this term is indicated by ≈) 2.718281829; see Appendix B for more details) while common logarithms are expressed as powers of 10 (Shennan, 1997). For example, 42 to the natural base ($\log_e$) is 3.738, which can be given by $e^{3.738}$, and 42 to the base 10 ($\log_{10}$) is 1.623, which can be given by $10^{1.623}$. Logarithms can be calculated using standard spreadsheet and GIS packages. Introductions to transformations are provided by Gregory (1968) and Shennan (1997).

The next section is concerned with analysing two or more variables simultaneously.

3.3 Multivariate statistics

The focus so far has been on methods for exploring single variables. In many applications, there is a need to consider how two or more variables are related to one another. For example, what is the relationship between altitude and snowfall? Does snowfall tend to be greater at high altitudes? This section makes use of the data given in Table 3.1. The first stage of an analysis may involve plotting one set of values against the other (this is called a scatter plot—an example is given in Figure 3.3) and such an analysis could be expanded using regression analysis, as detailed in this section.

Figure 3.3 indicates that small values of z (in this example, snowfall) tend to correspond to small values of y (elevation) and that large values of one tend to correspond

Table 3.1 Sample data set for illustrating regression.

Variable 1 (y) in m	Variable 2 (z) in cm
12	6
34	52
32	41
12	25
11	22
14	9
56	43
75	67
43	32

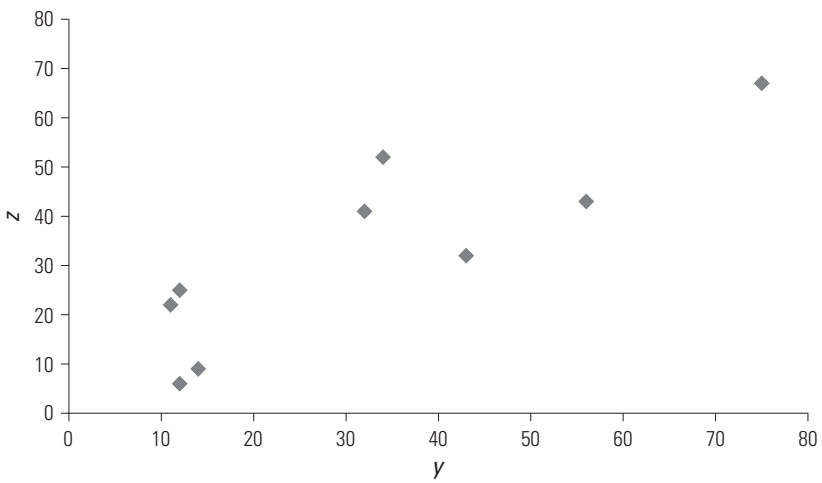

Figure 3.3 Scatter plot. Elevation (y) in metres (m) against snowfall (z) in centimetres (cm).

to large values of the other. Put another way, in general an increase in one corresponds to an increase in the other—it can be said that the variables are positively related to one another. If, as the values in one variable increase, the values of the other variable decrease (or vice versa) then the relationship is said to be negative. Analysis of relationships often proceeds using correlation and regression, which enable exploration of the nature of the relationship between two or more variables and the strength of the relationship between them. Correlation and regression are explored in this section.

In the case of two variables, regression is used to fit a line through the points on a scatter plot, this line being as close as possible to all the points according to some criterion. This is called the line of best fit. The line represents the trend in the data. If the variables are positively related, the line will be low with respect to the z-axis (representing small value) on the left of the graph and will increase diagonally from left to right. This would be the case for a line fitted to the plot in Figure 3.3. The correlation coefficient, r, provides a measure of the nature and strength of the relationship between variables. More specifically, it can be interpreted as indicating the degree to which points scatter around the regression line. Before detailing the measurement of correlation, the procedure for fitting a line to the scatter plot is detailed.

In this example, the variable y (elevation) is the independent variable while the variable z (snowfall) is the dependent variable. As well as allowing for exploration of the nature of the relationship between two variables, regression enables prediction of the values of dependent variables given values of independent variables. For example, if we have a raster map of elevation values across a region (i.e. we have values at all locations of interest) but only a few snowfall measurements, we could conduct regression by taking elevation values at the snowfall measurement locations and plotting these against the snowfall values. Once a line is fitted, the regression equation (indicating the form of the fitted line) can be used to predict snowfall values at locations where there are no snowfall measurements because the regression equation tells us what snowfall amounts to expect for any given value of elevation. This process is described below. The regression equation can be given by:

$$\hat{z}_i = \beta_0 + \beta_1 y_i \tag{3.5}$$

This indicates that the predicted value of z_i (with a prediction indicated by the hat on top of the letter)—the value given by the line of best fit—is obtained by adding β_0 to β_1 multiplied by y_i (in this example, the elevation value at location i). β is upper case Greek beta and these components are referred to as the beta coefficients. β_0 is called the intercept and is the point where the line crosses the vertical axis (representing the z variable in Figure 3.3). β_1 is called the slope coefficient. A negative value for the slope indicates a negative relationship and a positive value indicates a positive relationship. In this case, we know β_1 will be positive as the scatter plot shows that a higher elevation will correspond to a greater amount of snow. What is needed is a method to identify appropriate values of β_0 and β_1. Once we have these, we have

figures that tell us something about the nature of the relationship between variables. The least squares method is the most common approach to fitting a line to data and obtaining values of β_0 and β_1. This method minimizes the squared difference between the observed value (z_i, the measurement) and the value given by the line fitted with regression ($\hat{z}_i$):

$$\sum_{i=1}^{n}(z_i - \hat{z}_i)^2 \tag{3.6}$$

The following text describes how the intercept and slope are obtained through the ordinary least squares (OLS) method. The slope coefficient, β_1, is obtained from:

$$\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3.7}$$

The numerator gives the covariance between the independent and dependent values. The covariance is a measure of the degree to which two variables vary together and is the difference in one value from its mean multiplied by the difference in the second value from its mean. The covariances for each location are summed. The denominator is the sum of squared differences between the independent values and their mean.

The intercept, β_0, is given by:

$$\beta_0 = \frac{\sum_{i=1}^{n} z_i - \beta_1 \sum_{i=1}^{n} y_i}{n} = \overline{z} - \beta_1 \overline{y} \tag{3.8}$$

The values used in Table 3.1 are used to illustrate the regression procedure. Note that this sample is very small and in practice regression analyses should be based on much larger samples. However, this small sample allows direct illustration of the methods. This topic is discussed further in Section 3.4.

The slope is computed first. Initially, we compute the numerator of Equation 3.7:

$$\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})$$

We take each value of y and subtract the mean value of y; in turn we take each value of z and subtract the mean value of z. The difference between each y value and its mean and each z value and its mean is then multiplied together as shown in Table 3.2 (in the column headed $(y_i - \overline{y}) \times (z_i - \overline{z})$). This is done for all of the observations and the multiplied values are added together. The mean value of y is 32.11 and the mean value of z is 33.

The sum of the multiplied differences in Table 3.2 is 3092. The denominator of Equation 3.7, $\sum_{i=1}^{n}(y_i - \overline{y})^2$, is then used. In words, we take each value of y, subtract its mean, square this difference (see the column headed $(y_i - \overline{y})^2$ in Table 3.2), and add

Table 3.2 Variable 1 (y) and variable 2 (z), differences from their mean, differences multiplied, and the square of the differences from the mean of variable 1.

Variable 1 (y_i)	Variable 2 (z_i)	$(y_i - \overline{y})$	$(z_i - \overline{z})$	$(y_i - \overline{y}) \times (z_i - \overline{z})$	$(y_i - \overline{y})^2$
12	6	–20.11	–27.00	543.00	404.46
34	52	1.89	19.00	35.89	3.57
32	41	–0.11	8.00	–0.89	0.01
12	25	–20.11	–8.00	160.89	404.46
11	22	–21.11	–11.00	232.22	445.68
14	9	–18.11	–24.00	434.67	328.01
56	43	23.89	10.00	238.89	570.68
75	67	42.89	34.00	1458.22	1839.46
43	32	10.89	–1.00	–10.89	118.57

all of these squared differences together. The sum of these squared differences is 4114.89.

To compute the slope value, β_1, we divide the first summed value by the second:

$$\frac{3092}{4114.89} = 0.75142$$

The intercept, β_0, is then calculated:

$$\overline{z} - \beta_1 \overline{y} = 33 - (0.75142 \times 32.11) = 8.87190$$

In words, the intercept is given by the mean of the z values minus β_1 multiplied by the mean of the y values ($\beta_1 \overline{y}$ means that the two components are multiplied by one another and no multiplication symbol is needed). Note that a fairly large number of decimal places are used in the calculations to ensure that the manual calculations are close to those obtained using software packages.

The fitted line is shown in Figure 3.4. Regression is more conveniently conducted (i.e. values for β_0 and β_1 obtained) using matrix algebra, as outlined below, and such an approach is used in computer algorithms. The plot was generated using a spreadsheet package. Note that the intercept value in Figure 3.4 is slightly different to the figure given above, and the difference is due to rounding error in the calculations. The r^2 term in Figure 3.4 is the coefficient of determination and it is defined below.

Once we have values of β_0 (the intercept) and β_1 (the slope coefficient), we can make predictions. As an example, using the regression line shown in Figure 3.4, suppose we have a location with no snowfall measurement, but we know the elevation at that location is 43 units (e.g. metres). Following the regression equation

$$\hat{z}_i = \beta_0 + \beta_1 y_i$$

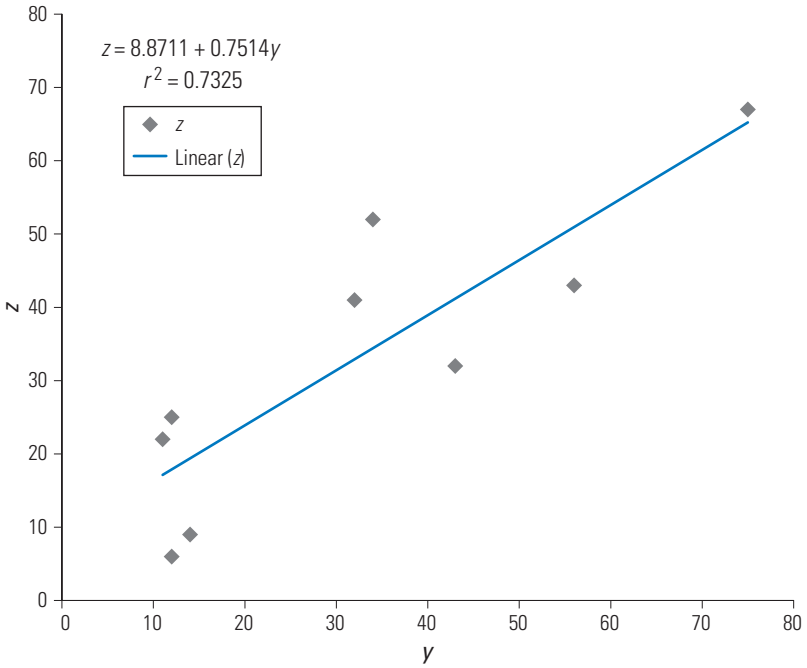

Figure 3.4 Scatter plot. Elevation (*y*) in metres (m) against snowfall (*z*) in centimetres (cm) with line of best fit.

we replace β_0 and β_1 with the values obtained for these previously and replace y_i (an elevation value in this case) with 43. This leads to:

$$\hat{z}_i = 8.87190 + (0.75142 \times 43) = 41.18296 \text{ cm}$$

For an elevation value of 43, therefore, the predicted value of snowfall (to three decimal places) is 41.183. This can be confirmed by looking at Figure 3.4 and drawing a line from the point corresponding to approximately 43 on the *y* (elevation) axis upwards to meet the line of best fit and then drawing a line from the point where the added line and the line of best fit meet across to the *z* (snowfall) axis. If the lines are accurately drawn, then a value of approximately 41 can be identified on the *z* (snowfall) axis.

The goodness of fit of a line of best fit can obtained by measuring the residuals, i.e. the difference between observed values and predicted values. As an example, Table 3.1 includes a *y* (elevation) value of 11 paired with a *z* (snowfall) value of 22. Using the approach outlined, the predicted value of snowfall for an elevation value of 11 is given by:

$$\hat{z}_i = 8.87190 + (0.75142 \times 11) = 17.13752 \text{ cm}$$

In this case the observed value is 22, and the predicted value to three decimal places is 17.138 so there is a difference (residual) of –4.862. In words, the regression model

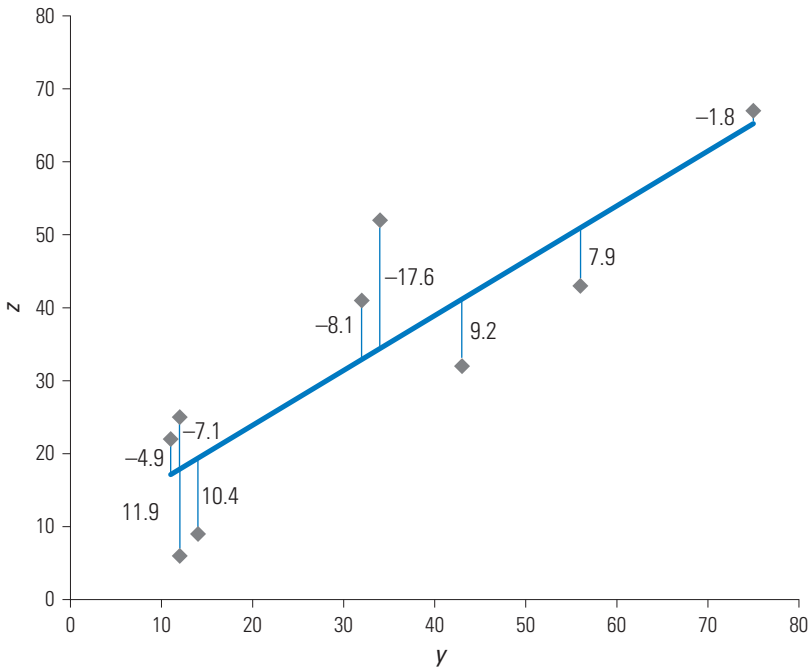

Figure 3.5 Scatter plot. Elevation (*y*) in metres (m) against snowfall (*z*) in centimetres (cm) showing residuals.

under-predicts the observed value by 4.862. Figure 3.5 shows residuals (to one decimal place) for the scatter plot in Figure 3.4.

Examining the residuals, perhaps by mapping them, may prove illuminating and may help to highlight regions with unusual characteristics. One goal of regression is to minimize the squared residuals while another is to minimize clustering in the values of the mapped residuals (see Sections 4.8 and 8.2 for discussions on a related topic).

Before discussing measures of goodness of fit, some grounding is briefly given in more efficient means for obtaining regression coefficients than was provided above. This background is necessary to enable readers to make the most of the descriptions of local regression procedures that come later. The following text introduces the idea of matrices and matrix multiplication—concepts essential to the worked example. Another key topic, inversion, is outlined in Appendix E.

Using matrix notation, the ordinary least squares (OLS) regression (the standard method of finding regression coefficients as outlined above) coefficients can be obtained using:

$$\beta = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{z} \tag{3.9}$$

The upper case bold letters indicate a set of values that can be arranged in a rectangle with at least two rows and columns. The lower case bold letter indicates the case

with only one column; this latter type of matrix is called a vector. The following are examples for the case of two variables and five cases (observations) of each:

$$\mathbf{Y} = \begin{bmatrix} 1 & y_1 \\ 1 & y_2 \\ 1 & y_3 \\ 1 & y_4 \\ 1 & y_5 \end{bmatrix}, \qquad \mathbf{Y}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ y_1 & y_2 & y_3 & y_4 & y_5 \end{bmatrix}, \qquad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix}$$

$\mathbf{Y}^T$ indicates the transpose of $\mathbf{Y}$—the flipped version of the original matrix (values in the left-hand column of $\mathbf{Y}$ are the top row in $\mathbf{Y}^T$ and values in the right-hand column of $\mathbf{Y}$ are the bottom row in $\mathbf{Y}^T$). The superscript −1 indicates the inverse, and this is explained in Appendix E. The '1' values for each entry in $\mathbf{Y}$ indicate that we are fitting a constant (intercept). In this example, we have five 1s, and five values of each of the independent (y) and the dependent (z) variables. Given Equation 3.9 and a knowledge of matrix algebra it is possible to find the regression coefficients for any number of independent variables. Computers make use of matrices, and so some understanding of how to use such methods is useful. Appendix E shows how Equation 3.9 is solved.

The solution to Equation 3.9 for the data presented above is given by (with the full working given in Appendix E):

$$\beta = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{z} = \begin{bmatrix} 0.36169 & -0.00780 \\ -0.00780 & 0.00024 \end{bmatrix} \times \begin{bmatrix} 297 \\ 12629 \end{bmatrix} = \begin{bmatrix} 8.871 \\ 0.751 \end{bmatrix}$$

where the calculations used to obtain the final values (the intercept and slope), shown in the matrix to the right-hand side above, are:

$$\beta_0 = (0.362 \times 297) + (-0.0078 \times 12629) = 8.871 \quad \text{(intercept)}$$
$$\beta_1 = (-0.0078 \times 297) + (0.00024 \times 12629) = 0.751 \quad \text{(slope)}$$

Note that the value of β_0 is smaller than the value obtained previously. This is due to rounding error (i.e. a different number of decimal places used in calculations).

The strength of the relationship between the variables can be measured using the correlation coefficient. The correlation coefficient, r, is given by:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \overline{z})^2}} \tag{3.10}$$

The numerator is the same as for the estimation of β_1 (see Equation 3.7). The denominator is simply the square root of the sum of squared differences between y

Table 3.3 Variable 1 (y) and variable 2 (z), squared differences from their mean, and summed values.

Variable 1 (y_i)	Variable 2 (z_i)	$(y_i - \overline{y})^2$	$(z_i - \overline{z})^2$
12	6	404.46	729.00
34	52	3.57	361.00
32	41	0.01	64.00
12	25	404.46	64.00
11	22	445.68	121.00
14	9	328.01	576.00
56	43	570.68	100.00
75	67	1839.46	1156.00
43	32	118.57	1.00
	Sum	4114.89	3172.00

and its mean multiplied by the square root of the sum of squared differences between z and its mean. These two sets of products are then multiplied together.

We have already seen the differences between y and z and their respective means in Table 3.2. The squared differences and their sums are given in Table 3.3 (note that the values in column 3 are the same as those in the final column of Table 3.2).

Given the values in Table 3.3, the numerator is obtained from:

$$\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z}) = 3092 \quad \text{(as calculated above)}$$

and the denominator is obtained from:

$$\begin{aligned}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \overline{z})^2} &= \sqrt{4114.89} \times \sqrt{3172.00} \\ &= 64.1474 \times 56.3205 \\ &= 3612.8136\end{aligned}$$

The correlation coefficient is then given by:

$$r = \frac{3092}{3612.8136} = 0.8558$$

The r value is positive and indicates, as the slope value, that the variables are positively related: as the value in one variable increases so does the value of the other. If the r value was negative, this would indicate negative correlation: as the value in one variable decreased, the value of the other would increase. Possible values of r range from −1 (indicating perfect negative correlation) to +1 (indicating perfect positive correlation).

The coefficient of determination is often used to indicate the goodness of fit, which is simply the squared correlation coefficient and is given by r^2. In our example (given

three decimal places), $r^2 = 0.856^2 = 0.732$. In this case, the coefficient of determination indicates that 73.2% of the variation in the data can be explained by the line of best fit. In words, the model represents the relationship quite well. So, an r^2 value close to zero indicates that the line (model) is a poor fit, while an r^2 value close to one indicates that the model is a good fit. Of course, where the relationship is non-linear (e.g. the scatter plot shows, along the horizontal axis, large values followed by small values, followed by large values, in a 'V' shape) then r (and r^2) may be close to zero and the scatter plot plays a key role in interpretation (Rogerson, 2006). The regression and correlation examples in this section are based on two variables (the dependent and an independent variable). Regression can easily be expanded to include more than one independent variable, thus allowing the assessment of the interrelationships between several variables simultaneously. In the case of more than one independent variable, upper case characters are used for the correlation coefficient and the coefficient of determination, thus R and R^2.

Some forms of data (e.g. nominal or categorical variables) should not be analysed directly using the methods outlined above. Alternative approaches are available in the case of values that are constrained to be whole numbers. Percentages and proportions should first be transformed before their analysis using standard statistical methods; Aitchison (1986) details some appropriate methods.

The topic of the following section is inferential statistics (i.e. statistical methods for making inferences about a population from a sample as opposed to descriptive statistics, which have been the focus in this section) and significance testing (e.g. testing for the significance of the differences between groups). As an example, it is standard practice to ascertain the significance of regression coefficients or the correlation coefficient, and the testing of the latter is outlined below.

3.4 Inferential statistics

In this section, the focus is on statistical methods for making inferences about a population from a sample as opposed to statistics which simply summarize a sample. Two common tasks in inferential statistics contexts are (1) to consider the likelihood that a statement about a given parameter (e.g. the mean or standard deviation) is true given the available data and (2) to estimate the parameters (Brunsdon, 2008). The first of these relates to the concept of hypothesis testing while in the second the confidence interval is central.

A common objective in statistical inference is to compare sets of samples and assess the degree of difference between the samples. In other words, we may be interested in assessing the probability that two samples come from different populations. Comparison of samples is based on tests of significance. In words, we test the significance of the difference between two (or more) samples to assess if the difference between them is likely to be 'real' in some sense. Questions of differences between samples are usually phrased in terms of a null hypothesis and the alternative hypothesis, indicated

by H_0 and H_1, respectively. In simple terms, H_0 could be the hypothesis that there is no significant difference and H_1 the hypothesis that there is a significant difference. The significance level, α (lower case Greek alpha), can be defined as the probability (defined below) that the null hypothesis is correct (Ebdon, 1985). If some outcome is statistically significant then it is regarded as unlikely to have occurred by chance. Hypothesis testing is therefore the procedure of rejecting or accepting the null hypothesis. The possible errors associated with this process can be defined as follows:

type I error: reject the null hypothesis when it is true

type II error: accept the null hypothesis when it is false

The emphasis is on avoiding type I errors on the grounds that accepting the null hypothesis when it is false is likely to be less damaging than wrongly accepting the hypothesis that there are real differences.

It is useful to have some indication of how likely it is that a sample is representative of a population. For example, if we have the mean of the salaries of a set of individuals how confident can we be that this is representative of the population as a whole in the study area? This question can be approached by computing the standard error of the mean:

$$SE_{\mu} = \frac{s}{\sqrt{n}} \tag{3.11}$$

where μ is the mean average, s is the sample standard deviation, both as defined previously, and n is the number of observations in the sample. The standard error determines what is known as the confidence interval and a small standard error gives greater confidence that the sample mean is close to the population mean. The next stage of the discussion requires some knowledge of the concept of probability. A probability can be defined as the likelihood of a given outcome and is expressed as a fraction of 1. If a given event occurs 70 times out of 100 then the probability of the event occurring is expressed as 0.7. Given Equation 3.11, for a normal distribution, there is a 0.682 probability that the population mean is within one standard error of the mean and a 0.954 probability that the population mean is within two standard errors of the mean (see Section 3.2):

$$SE_{\mu} = \frac{4.14}{\sqrt{64}} = \frac{4.14}{8} = 0.518$$

In words, there is a 0.682 probability that the population mean is 37 ± 0.518 (or 36.482 to 37.518). By widening the confidence interval, we can say that there is a 0.954 probability that the population mean is 37 ± 1.035 (or 35.965 to 38.035).

As well as assessing confidence in the mean of a single sample, there is often a desire to assess differences between two different sample means. One way of doing this is to use the *t*-statistic (introduced in a specific context below). Introductions are provided by Shennan (1997), Ebdon (1985), and Rogerson (2006). When the desire is to compare

multiple categories, analysis of variance (ANOVA) represents the standard framework. The ANOVA test can be used to compare variation within data columns to variation between data columns. In words, given a null hypothesis that a set of population means are equal, we would reject the null hypothesis if the variation between the means of each group is significantly greater than the variation within the data columns (Rogerson, 2006).

Section 3.3 discussed correlation and regression. If it is assumed that the distributions of the variables are normal and the observations of each variable are independent of one another, then a test of the significance for the correlation coefficient, r, may be conduced using the t-statistic (Rogerson, 2006):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{3.12}$$

Given the small example in Section 3.3, where r was 0.856 and n was 9, this gives:

$$t = \frac{0.856\sqrt{9-2}}{\sqrt{1-0.856^2}} = \frac{0.856\sqrt{9-2}}{\sqrt{1-0.856^2}} = \frac{2.265}{0.517} = 4.381$$

Before assessing this result, a little explanatory text is required. Hypotheses can be one-sided or two-sided. In the former case, a test is conducted to assess if the true value is above *or* below (but not both) a given value. In the latter case, the test considers if the true value is *either* side of a given value. This example relates to a two-sided (or two-tailed) test.

Given these values, examining a t-table (a table allowing identification of the significance level associated with a t value given a particular number of degrees of freedom (see, for example, Shennan, 1997; Ebdon, 1985)) indicates that, for a two-tailed test with $\alpha = 0.05$ (i.e. the 5% significance level, commonly used as a benchmark significance level) and with seven degrees of freedom (there are nine observations and $n-2=9-2=7$) the critical value of t is 2.365. Since 4.381 is greater than that value, the correlation coefficient can be said to be not equal to zero and the null hypothesis is rejected. An alternative is to use one of the many web-based t-test calculators. Of course, standard software packages will do the calculations for you in any case. It is important to take into account the sample size when assessing the correlation coefficient (or other coefficient) as, while the correlation coefficient may suggest a strong relationship between variables, there may in fact be little confidence in the results if the sample size is small.

3.5 Statistics and spatial data

The focus of this book is the analysis of *spatial* data. Spatial data cannot be blindly treated in the same way as data that are not located spatially (i.e. aspatial data).

One key problem relates to statistical inference, as discussed above. One of the assumptions of standard significance tests is independence in the observations of the samples. In simple terms, this means that the value of one observation should not be affected by the value of another observation. As will be shown in Section 4.8, values of observations of spatial variables are often very similar to the values at neighbouring locations. Indeed, the fact that this is so frequently the case provides the basis of many spatial data analysis approaches. In the present context, the problem with this is that significance levels tend to be inflated as the effective number of degrees of freedom is reduced—that is, if the observations are spatially dependent then there are effectively fewer independent observations. In words, use of standard significance tests with spatial data is problematic and solutions may not be straightforward. Rogerson (2006) provides a detailed discussion of this topic.

Summary

This chapter provides a summary of some key ideas and methods in descriptive and inferential statistical analysis. Many methods for the analysis of spatial data have their origins in standard aspatial statistical methods, and knowledge of some key methods and issues in statistical analysis is essential for the developing spatial analyst, who will need to make frequent use of standard statistical approaches in the initial exploration of spatial data sets. Many of the themes introduced here will be revisited in later chapters, but with adaptations reflecting the focus of this book on spatial data.

Further reading

The book by **Rowntree (2000)** provides an excellent introduction to some important statistical concepts. Various books provide introductions to statistics for geographers and the book by **Rogerson (2006)** is a good example. The books by **Kitchin and Tate (2000)** and **O'Sullivan and Unwin (2002)** also include relevant introductory material. **Rowntree (2000)** and **Rogerson (2006)** include very clear introductions to some key concepts in inferential statistics. **Brunsdon (2008)** provides an introduction to some key concepts in statistical inference and discusses some issues related to inference in a spatial context. Section 7.3.1 illustrates the chi-square (χ^2) test, a commonly used statistical hypothesis test.

➜ The next chapter explores some key concepts in *spatial* data analysis and is the third of the three chapters introducing key concepts that provide foundations for the rest of the book.