**Appendix**

**Dataset**

This table gives the simulated dataset that is used for much of this chapter (source: original). Each row represents a site, for which the attributes in the columns were measured. Units are mm per year for rainfall, °C for mean temperature and grams per year for plant growth. Fertilizer is recorded as applied to the plants or not, and light levels are recorded as low (shade), medium (semi-shade) or high (no shade).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rainfall** | **Temperature** | **pH** | **Fertilizer** | **Light** | **Plant growth** |
| 696 | 17.2 | 8.3 | no | Low | 37.9 |
| 686 | 17.4 | 5.9 | no | Low | 46.2 |
| 803 | 19.7 | 4.2 | no | Low | 43.5 |
| 609 | 14.6 | 5.6 | no | Low | 30.3 |
| 728 | 16.0 | 5.7 | no | Low | 41.1 |
| 980 | 10.4 | 8.7 | no | Low | 19.2 |
| 976 | 17.7 | 7.1 | no | Low | 28.3 |
| 755 | 14.0 | 7.4 | no | Low | 46.7 |
| 811 | 22.9 | 6.0 | no | Low | 51.4 |
| 638 | 18.0 | 5.3 | no | low | 53.5 |
| 898 | 20.6 | 7.6 | no | medium | 64.0 |
| 955 | 16.9 | 7.0 | no | medium | 58.2 |
| 837 | 14.6 | 8.1 | no | medium | 43.5 |
| 595 | 12.3 | 4.8 | no | medium | 49.3 |
| 678 | 22.6 | 7.9 | no | medium | 55.3 |
| 899 | 15.0 | 6.4 | no | medium | 51.3 |
| 950 | 24.3 | 7.6 | no | medium | 57.8 |
| 533 | 10.7 | 7.9 | no | medium | 37.6 |
| 668 | 18.5 | 5.6 | no | medium | 59.6 |
| 573 | 11.8 | 7.2 | no | medium | 48.8 |
| 798 | 10.8 | 8.7 | no | high | 44.9 |
| 947 | 13.2 | 7.2 | no | high | 53.1 |
| 838 | 12.3 | 4.8 | no | high | 42.1 |
| 968 | 21.1 | 8.1 | no | high | 73.3 |
| 974 | 13.5 | 5.2 | no | high | 55.0 |
| 619 | 11.0 | 4.5 | no | high | 43.0 |
| 908 | 23.3 | 8.4 | no | high | 62.3 |
| 981 | 12.2 | 8.5 | no | high | 40.7 |
| 862 | 19.2 | 7.7 | no | high | 55.1 |
| 894 | 21.5 | 4.6 | no | high | 58.0 |
| 575 | 14.4 | 4.3 | yes | low | 98.3 |
| 543 | 15.7 | 5.3 | yes | low | 105.0 |
| 839 | 22.3 | 5.9 | yes | low | 128.3 |
| 906 | 21.5 | 4.6 | yes | low | 125.5 |
| 924 | 17.2 | 5.6 | yes | low | 114.4 |
| 718 | 10.8 | 8.0 | yes | low | 90.2 |
| 521 | 18.3 | 6.7 | yes | low | 96.8 |
| 624 | 21.1 | 4.3 | yes | low | 129.9 |
| 826 | 19.4 | 7.7 | yes | low | 117.4 |
| 803 | 20.1 | 8.4 | yes | low | 126.5 |
| 533 | 23.4 | 8.2 | yes | medium | 144.7 |
| 880 | 14.4 | 6.1 | yes | medium | 128.2 |
| 923 | 22.9 | 6.8 | yes | medium | 166.5 |
| 692 | 16.0 | 5.3 | yes | medium | 145.9 |
| 975 | 24.6 | 4.7 | yes | medium | 157.5 |
| 764 | 18.5 | 4.8 | yes | medium | 145.1 |
| 520 | 19.8 | 8.7 | yes | medium | 137.5 |
| 518 | 20.9 | 4.6 | yes | medium | 126.7 |
| 568 | 12.9 | 7.3 | yes | medium | 120.4 |
| 514 | 13.4 | 6.0 | yes | medium | 130.0 |
| 957 | 23.6 | 5.8 | yes | high | 165.0 |
| 837 | 20.8 | 6.7 | yes | high | 172.1 |
| 656 | 19.1 | 6.9 | yes | high | 148.5 |
| 810 | 23.7 | 8.9 | yes | high | 156.8 |
| 802 | 21.8 | 8.5 | yes | high | 174.2 |
| 873 | 11.8 | 6.2 | yes | high | 144.0 |
| 537 | 20.1 | 7.5 | yes | high | 133.2 |
| 600 | 19.3 | 6.2 | yes | high | 156.1 |
| 930 | 20.3 | 6.9 | yes | high | 161.6 |
| 587 | 12.6 | 4.2 | yes | high | 125.4 |

**Principles important for exploring, analysing and presenting data**

*Types of data*

Most standard statistical textbooks have chapters explaining the differences between types of data (e.g. Ebdon, 1985). Various labels are attached to these types but the primary distinction is between categorical and continuous data. Continuous data are those that are measured on a continuous scale and so may have any value, including fractions. The most important thing about continuous data is the relational element: a meaningful progression as the numbers increase. For example, a length of 5 mm is twice a length of 2.5 mm (an attribute of ‘ratio data’) and 6 mm is less than 11 mm (an attribute of the more general ‘interval data’ type). This relational property means that continuous data differ fundamentally from categorical data, in which each value signifies membership of a particular group, such as ‘teacher’ or ‘student’ for the variable ‘occupation’, or ‘Thames’ or ‘Ganges’ for the variable ‘river’.

In Table 1, notice how there are no unique values for the categorical variables: members of the same group have the same value (e.g. ‘lawyer’), by definition. In contrast, most or all of the values for a continuous variable are unique (depending on the level of precision used for recording the data). It is important to realise that categorical variables such as sex or occupation *are* variables – there is a value assigned to each unit – they just differ in data type from continuous variables. The value assigned can be text (e.g. ‘female’) or numeric (e.g. a value of 1 used as code for ‘female’). Statistical packages on computers tend to use the numeric form (internally at least), but some (such as SPSS) allow both numbers and text labels to be used simultaneously. Note that, if a categorical variable is recorded as numbers, these numbers are not relational like they are in a continuous variable: ‘female’ could be recorded as 1 and ‘male’ as 2, but it makes no sense to say that ‘male’ is one more than (or indeed twice) ‘female’!

Continuous data can be reduced to groups which are categorical but ordered, so retain a relational element. For example, fertilizer application could be recorded as ‘low’, ‘medium’ or ‘high’ rather than kg/ha. This loses information and so is not generally recommended. Further, it is important to realize that, while exact measurements can easily be converted to this form, the reverse is not true: data collected in this form cannot be converted into exact measurements. What exact amount, in kg/ha, corresponds to a fertilizer application recorded as ‘high’? Many of the most powerful types of statistical analysis, such as regression, require data in continuous form and so, as a general rule, it is better to record exact measurements when collecting data. There are three main exceptions to this rule. The first is trivial: some variables can only sensibly be measured as categories. In Table 1, sex and occupation are examples of this. Secondly, there are times when the apparent precision of an exact measurement does not reflect the accuracy of that measurement. A common example is the recording of people’s ages in questionnaire surveys: in many situations, it has been shown that people have a greater propensity to lie about their age when asked to give a figure, than when asked which age category they belong to. In such cases, accuracy can actually be improved by the loss of apparent precision (the gain is in the reduced bias). The third major exception to the rule is when a better overall impression of the sampled population is likely to be gained by collecting a lot of categorical, rather than a few continuous data. For example, recording exact percentage vegetation cover in quadrats is immensely time consuming and difficult. One alternative is to record a rough estimate, done quickly by eye. While this gives an ‘exact’ value its accuracy is very dubious. Another alternative is to use a categorical scale, such as recording the vegetation as ‘<1%’, ‘1-4%’, ‘4-10%’, ‘10-25%’, etc. (Many such methods exist, two of the most well-known being the Domin and the Braun-Blanquet scales.) This kind of method allows the rapid collection of data without producing values that are misleading in their apparent precision.

Other data types include discreet data (or integers), which are like continuous data but cannot include fractional numbers, and categories that can be ordered (as in vegetation cover measured using the Domin scale) – so that the difference between a ‘1’ and a ‘2’ is more meaningful than for purely categorical data. Some data are bounded: values cannot lie above a certain number and/or below another. A good example is percentage data (for a fuller discussion of types of data, see standard statistical textbooks, e.g. Ebdon, 1985).

*What to check for normality: the distribution of errors, NOT the distribution of the raw data*

A key assumption of most parametric statistical techniques is that the distribution of sample statistics can be assumed to conform to one of the known theoretical distributions. Because the exact probability function of each of these theoretical distributions is known, all sorts of probability statements can be made – *as long as the assumption holds*. Many of the more familiar statistical techniques (such as regression, ANOVA and t-tests) assume that the data conform to the normal distribution. Hence the value of tools like the graphs shown in Figure 5, and also numerical statistical tests for normality (Sokal and Rohlf, 1995), which we can use to establish the validity of this key assumption. But herein lies a common misunderstanding.

What should we test for normality? More formally, the assumption of the relevant statistical techniques is that the errors (or residuals) are normally distributed. Normal distribution of the raw data for a response variable will often mean normal distribution of residuals from a statistical model and, for this reason, many people examine the distribution of variables prior to statistical analysis. However, results from such an exercise should be treated with caution, as errors can be normally distributed even if the response variable is not. [See main text for explanation of terms like ‘response variable’ and ‘error’.]

In Figure 5b, the distribution of the data is clearly not normal. It is bimodal (i.e. two modes, or concentrations of values along the horizontal-axis scale). When performing statistical modelling, many departures from normality of errors can be cured by transforming variables, but this may not be possible with bimodal errors. Fortunately, however, bimodality of distribution of the raw data of a response variable is usually due to underlying structure in the data, and if that can be accounted for the bimodality disappears. For example, the distribution of people’s heights is typically bimodal because men tend to be taller than women; but when men and women are analysed separately, the heights tend to be normally distributed. This simple example neatly illustrates the folly of only examining the distribution of the raw data. The data in Figure 5b–e represent a similar scenario.

*Evaluating statistical models*

Whenever we consider a model (such as a regression), we should ask searching questions about what its purpose is ‒ is it for prediction or merely description, for example ‒ and about exactly what it tells us. We should also think about what alternatives there might be. Would a different model work better ‒ perhaps a non-linear rather than a linear relationship, or vice-versa? How robust is the model? Good practice in both mathematical and statistical modelling involves sensitivity analysis, in which the effects of changes in assumptions or input variables are evaluated to inform us about this.

*Statistics – an Intuitive Introduction (‘SII’)*

If you wish to teach yourself the basics of classical statistics, including a range of standard presentation methods, there is a free learning package available at www.nottingham.ac.uk/toolkits/play\_244 or see http://www.nottingham.ac.uk/~lgzwww/prezi.html for an enhanced menu. This walks you through the basics in an interactive way, allowing you to build up your knowledge and understanding at your own pace, and in as much depth as you wish. It contains practice exercises using randomly generated data, so there is an infinite supply of questions (with answers).

**References**

Ebdon, D.S. (1985) *Statistics in Geography* (2nd edn). Oxford: Blackwell.

Sokal, R.R and Rohlf, F.J. (1995) *Biometr*y. New York: W.H. Freeman.