CHAPTER 1

Introduction

It's all around us.

In this book we will be examining how statistical principles and methods can be used to study environmental problems. Our concern will be directed to:

- probabilistic, stochastic and statistical models;
- data collection, monitoring and representation;
- drawing inferences about important characteristics of the problem;
- using statistical methods to analyse data and to aid policy and action.

The principles and methods will be applicable to the complete range of environmental issues (including pollution, conservation, management and control, standards, sampling and monitoring) across all fields of interest and concern (including air and water quality, forestry, radiation, climate, food, noise, soil condition, fisheries, and environmental standards). Correspondingly, the probabilistic and statistical tools will have to be wide-ranging. *Inter alia*, we will consider extreme processes, stimulus response methodology, linear and generalized linear models, sampling principles and methods, time series, spatial models and methods, and, where appropriate within these themes, give attention to appropriate multivariate techniques and to design considerations including, for example, designed experiments.

Any models or methods applicable to situations involving uncertainty and variability will be relevant in one guise or another to the study and interpretation of environmental problems and will thus be part of the armoury of *environmental statistics* or *environmetrics*. Environmental statistics is a vast subject. In an article in the journal *Environmetrics*, Hunter (1994) remarked: 'Measuring the environment is an awesome challenge, there are so many things to measure, and at so many times and places'. But, however awesome, it must be faced! The recently published four-volume *Encyclopaedia of Environmetrics* (El-Shaarawi and Piegorsch, 2002) bears witness to the vast coverage of our theme and to its widespread following.

Environmental Statistics V. Barnett
 ISBN: 0-471-48971-9 (HB)

1.1 TOMORROW IS TOO LATE!

As we enter the new millennium the world is in crisis – in so many respects we are placing our environment at risk and not reacting urgently enough to reverse the effects. Harrison (1992) gives some graphic illustrations (see also Barnett, 1997):

- The average European deposits in a lifetime a monument of waste amounting to about 1000 times body weight; the average North American achieves four times this.
- Sea-floor sediment deposits around the UK average 2000 items of plastic debris per square metre.
- Over their lifetime, each person in the Western world is responsible for carbon dioxide emissions with carbon content on average 3500 times the person's body weight.

The problems of acid rain, accumulation of greenhouse gases, climate change, deforestation, disposal of nuclear waste products, nitrate leaching, particulate emissions from diesel fuel, polluted streams and rivers, etc., have long been crying out for attention. Ecological concerns and commercial imperatives sometimes clash when we try to deal with the serious environmental issues. Different countries show different degrees of resolve to bring matters under control; carbon emission is a case in point, with acclaimed wide differences of attitude and practice between, for example, the United States and the European Union. Environmental scientists, and specialists from a wide range of disciplines, are immersed in efforts to try to understand and resolve the many environmental problems we face.

Playing a major role in these matters are the statisticians, who are uniquely placed to represent the issues of uncertainty and variation inevitably found in all environmental issues. This is vital to the formulation of models and to the development of specific statistical methods for understanding and handling such problems.

1.2 ENVIRONMENTAL STATISTICS

Environmental statistics is a branch of statistics which has developed rapidly over the past 10–15 years, in response to an increasing concern among individuals, organizations and governments for protecting the environment. It differs from other applications topics (e.g. industrial statistics, medical statistics) in the very wide range of emphases, models and methods needed to encompass such broad fields as conservation, pollution evaluation and control, monitoring of ecosystems, management of resources, climate change, the greenhouse effect, forests, fisheries, agriculture and food. It is also placing demands on the statisticians to develop new approaches (e.g. to spatial-temporal modelling) or new methods (e.g. for sampling when observations are expensive or elusive

or when we have specific information to take into account) as well as to adapt the whole range of existing statistical methodology to the challenges of the new environmental fields of application.

Environmental statistics is indeed becoming a major, high-profile, identified theme in most of the countries where statistical analysis and research are constantly advancing our understanding of the world we live in. Its growing prominence is evident in a wide range of relevant emphases throughout the world.

Almost all major international statistical or statistically related conferences now inevitably include sessions on environmental statistics. The International Environmetrics Society (TIES), which originated in Canada, has, over more than a decade, held in excess of ten international conferences on environmental statistics and has promoted the new journal *Environmetrics* (published by John Wiley & Sons, Ltd). The SPRUCE organization, established in 1990, is concerned with *Statistics in Public Resources and Utilities, and in Care of the Environment* and has also held major international conferences. Four resulting volumes have appeared under the title *Statistics for the Environment* (Barnett and Turkman, 1993, 1994, 1997; Barnett *et al*., 1999). A further volume on *Quantitative Methods for Current Environmental Issues* covers the joint SPRUCE–TIES conference in Sheffield, UK, held in September 2000 (Anderson *et al*., 2002).

Other expressions of concern for environmental statistics are found in the growing involvement of national statistical societies, such as the Royal Statistical Society in the UK and the American Statistical Association, in featuring the subject in their journals and in their organizational structure. Specific organizations such as the Center for Statistical Ecology and Environmental Statistics at Penn State University, US, and the broader-based US Environmental Protection Agency (USEPA) are expanding their work in environmental statistics. Other nations also express commitment to the quantitative study of the environment through bodies concerned with environmental protection, with environmental change networks and with governmental controls and standards on environmental emissions and effects. Many universities throughout the world are identifying environmental statistics within their portfolios of applications in statistical research, education and training.

Of course, concern for quantitative study of environmental issues is not a new thrust. This is evidenced by the many individuals and organizations that have for a long time been involved in all (including the statistical) aspects of monitoring, investigating and proposing policy in this area. These include health and safety organizations; standards bodies; research institutes; water and river authorities; meteorological organizations; fisheries protection agencies; risk, pollution, regulation and control concerns, and so on.

Such bodies are demanding more and more provision of sound statistical data, knowledge and methods at all levels (from basic data collection and sampling to specific methodological and analytic procedures). The statistician is of course ideally placed to represent the issues of uncertainty and variation inevitably found in all environmental problems. An interesting case in point

was in relation to the representation of uncertainty and variation in the setting of environmental pollution standards in the 1997 UK Royal Commission on Environmental Pollution study (see Royal Commission on Environmental Pollution, 1998; Barnett and O'Hagan, 1997).

Environmental statistics is thus taking its place besides other directed specialities: medical statistics, econometrics, industrial statistics, psychometrics, etc. It is identifying clear fields of application, such as pollution, utilities, quality of life, radiation hazard, climate change, resource management, and standards. All areas of statistical modelling and methodology arise in environmental studies, but particular challenges exist in certain areas such as official statistics, spatial and temporal modelling and sampling. Environmentally concerned statisticians must be pleased to note the growing public and political acceptance of their role in the environmental debate.

Many areas of statistical methodology and modelling find application in environmental problems. Some notable emphases and needs are as follows:

- The study of extremes, outliers and robust inference methods is relevant to so many fields of inquiry, none more so than environmental issues.
- Particular modern sampling methods have special relevance and potential in many fields of environmental study; they are important in monitoring and in standard-setting. For example, ranked-set sampling aims for high-efficiency inference, where observational data are expensive, by exploiting associated (concomitant, often 'expert-opinion') information to spread sample coverage. Composite sampling seeks to identify rare conditions and form related inferences again where sampling is costly and where sensitivity issues arise, whilst adaptive sampling for elusive outcomes and rare events modifies the sampling scheme progressively as the sample is collected.
- Other topics such as size-biasing, transect sampling and capture–recapture also find wide application in environmental studies.
- Linear and generalized linear models play a central role in statistical investigations across all areas of environmental application. Further developments are needed, particularly in relation to multivariate correlated data, random dependent variables, extreme values, outliers, complex error structure, etc., with special relevance to environmental, economic and social issues.
- Risk evaluation and uncertainty analysis are modern thrusts which still need more careful definition and fuller investigation to formalise their pivotal roles and to elucidate distinctions with conventional statistical concepts and methods (see Barnett and O'Hagan, 1997; Barnett, 2002d). Applications of special relevance occur across the environmental spectrum.
- Temporal and spatial models have clear and ubiquitous relevance; all processes vary in time and space. Time-series methods have been widely applied and developed for environmental problems but more research is needed on non-stationary and multivariate structures, on outliers and on non-parametric approaches. Spatial methods need further major research development, including concern for the highly correlated and multivariate

base of most applications fields. Conjoint spatial-temporal models and methods are not well developed and are ripe for major advances in research.

In this book we will seek to review and represent the wide range of applications and of statistical topics in environmental statistics. This is facilitated by dividing the coverage into a number of thematic parts as follows:

Part I Extremal stresses: extremes, outliers, robustness (Chapters 2 and 3);
Part II Collecting environmental data: sampling and monitoring (Chapters 4, 5 and 6);
Part III Examining environmental effects: stimulus–response relationships (Chapters 7 and 8);
Part IV Standards and regulations (Chapter 9);
Part V A many-dimensional environment: spatial and temporal processes (Chapters 10 and 11).

Each part is, as indicated, divided into separate chapters covering different appropriate aspects of the respective theme.

1.3 SOME EXAMPLES

We will start our study of environmental statistics by considering briefly some practical examples, from different fields, which also illustrate various models and methods which will be developed more formally and in more detail as the book progresses.

1.3.1 'Getting it all together'

Collecting data in an effective and efficient manner is of central importance in studying environmental problems. Often we need to identify those members of a population who possess some rare characteristic or condition or to estimate the proportion of such members in the population. Sometimes the condition is of a 'sensitive' form, and individuals may be loath to reveal it. Alternatively, it may be costly or difficult to assess each member separately.

One possibility might be to obtain material or information from a large group of individuals, to mix it all together and to make a single assessment for the group as a whole. This assessment will reveal the condition if any one of the group has the condition. If it does not show up in our single test we know that all members are free of the condition. A single test may clear 1000 individuals!

This is the principle behind what is known as *composite sampling* (or composite testing). It is also known as *aggregate sampling* or in some contexts as *grab sampling*. For example, we might use a dredger to grab a large sample of soil deposit from a river bed and conduct a single test to show there is no contamination.

Of course, our composite sample might show the condition to be present. We then know nothing about which, or how many, individuals are affected. But that is another matter to which we will return later. If the condition is rare, the single composite sample will often be 'clear' and we will be able to clear all members of the sample with a single test.

Such an approach is not new – early examples of such group testing were concerned with the prevalence of insects carrying a plant virus (Watson, 1936) and of testing US servicemen for syphilis in the Second World War (Dorfman, 1943). An informative elementary review of composite sampling applications is given by USEPA (1997).

So how does this method operate? Typically a (usually large) number of random samples of individuals are chosen from the population. The material collected from each member of a sample is pooled, and a single test carried out to see if the condition is present or absent; for example, blood samples of patients might be mixed together and tested for the presence of the HIV virus.

Thus, suppose that our observational samples are of sizes $n_1, n_2, n_3, \ldots$ (often all n_i are equal) and that our corresponding composite test outcomes are 0, 0, 1, ... (where 0 means negative and 1 means positive). From these data, we can develop an estimator $\tilde{p}$ of the crucial parameter

$$p = P \text{ (individual has the condition)}$$

which expresses the rate of incidence of the condition in the population at large. Further, we will be able to derive the statistical properties of the estimator (i.e. to examine whether it is biased, to determine what is its variance or mean square error, etc.).

An interesting situation arises when, rather than estimating p, our interest is in identifying which specific members of a sample actually have the condition. If the test does not show the condition, all members are free of it. But if it is present, we must then do more testing to find out which members have the condition. Different strategies are possible. Suppose we start with n individuals. The most obvious possibility is *full retesting* (Figure 1.1). Here we do the overall composite test and if it is positive we then proceed to retest each individual, so

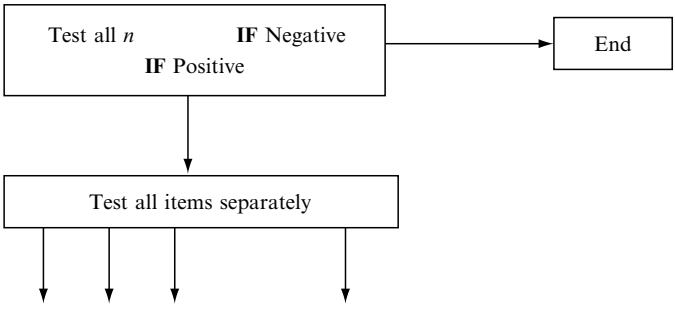

Figure 1.1 Full retesting.

that we either do one test (if the first test result is negative) or $n+1$ tests (otherwise) to uncover the complete situation. If the condition is rare (p is small) the expected number of tests will be just a little larger than 1, compared with the n needed if we just tested all individuals separately at the outset. For example, if $n = 20$ and $p = 0.0005$ the expected number of tests turns out to be just 1.2!

An alternative approach (illustrated in Figure 1.2) is *group retesting*. Here the sample group of size n is divided into k subgroups of sizes $n_1, n_2, \ldots, n_k$ if the first overall composite test is positive, and each of the subgroups is treated as a second-stage composite sample. Each subgroup is then tested as for full retesting and the process terminates.

Different strategies for choice of the number k and sizes $n_1, n_2, \ldots, n_k$ of subgroups have been considered, yielding crucial differences in efficiency of identification of the 'positives' in the overall sample. It is interesting to examine this by trying out different choices of k and $n_1, n_2, \ldots, n_k$.

A special, useful modification of group retesting is *cascading*, where we adopt a hierarchical approach, dividing each positive group or subgroup into precisely two parts and continuing testing until all positives have been identified. Figure 1.3 shows how this might operate.

These three strategies provide much scope for trying to find effective means of identifying the positive individuals – and also for some interesting combinatorial probability theory calculations.

Composite sampling can also be used to estimate characteristics of quantitative variables in appropriate circumstances, such as the mean density of polluting organisms in a water supply, as we shall find in our later, more detailed discussion of the approach (Section 5.3). This intriguing method of composite sampling is just one example of the many modern sampling methods that are being used to obtain data on environmental matters.

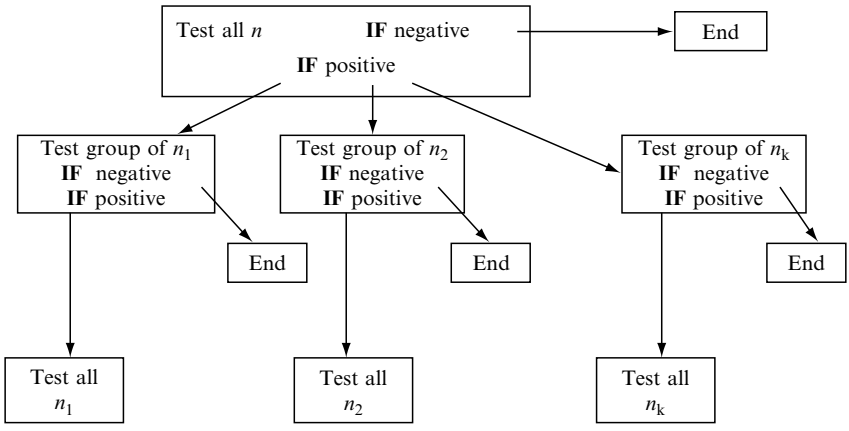

Figure 1.2 Group retesting.

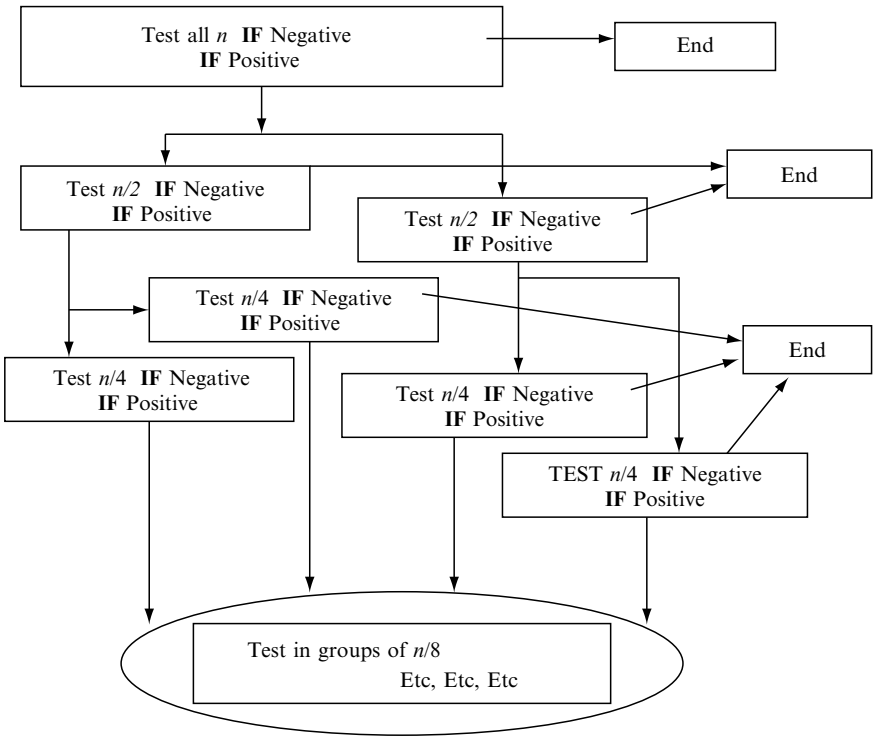

Figure 1.3 Cascading.

1.3.2 'In time and space'

Environmental effects involve uncertainty and need to be represented in terms of random variables. They frequently also exhibit systematic effects related to concomitant variables. For example, the extent of nitrate leaching through fertilized agricultural land may well depend on the concentration of recently applied nitrogenous fertilizer. Thus we will need to study relationships between variables. These may be expressed in simple linear regression model terms, or may require more complicated models such as the generalized linear model or speific non-linear models.

As an example, we might consider how the mid-day temperature on a specific day, at different sites in the UK, varies with the northings and eastings of the sites, which specify where they are located. Perhaps we might feel that the temperature will tend to be lower the more northerly the site (and perhaps the more easterly). This could be expressed in terms of a regression model with two regressor variables representing northing and easting. But what about altitude? This could also be influential.

In this example we see a variable which varies over space – *spatial models* will need to be a feature of our studies and will go beyond simple regression structures. Again, we have considered just a single day. Clearly the time of

the year, even the time of the day (if not just noon) will be crucial to the levels of temperature we might expect to encounter. So there will also be a temporal component, and we will need to model the temperature in terms which allow for the effect of time. *Time-series* methods are designed to do this, and we will examine their principles and procedures. The practical example discussed in Section 1.3.3 below also needs precisely such a spatial and temporal basis for its study and raises the prospect of whether we could contemplate a *joint spatial-temporal model*. This is a largely uncharted area.

Regression models are widely used for spatial variation (as are prediction or mapping procedures such as *kriging* – see Section 11.2), while time-series models are used represent time variation; but how are we to combine the spatial and temporal components?

For the moment we notice that a linear (regression) model for the mid-day temperature which expresses the temperature as a linear combination of space and time variables could in principle provide such a *joint model*, but it would have clear limitations.

Consider another atmospheric variable. It is important for various purposes to monitor the amount of ground-level ozone as an environmental health indicator. But where should we measure it (on the roof of the local railway station?) and when (Tuesday, at noon?). In fact we want to know (and to economically represent) how the ozone level is varying from place to place over a defined region and from time to time over a specified time period, as we did for the temperature.

That is, we need realistic models for $Z(t, \mathbf{s})$ (ozone level at time t and location $\mathbf{s}$) which ideally would reflect levels and intercorrelations from time to time and place to place. We need then to develop statistical methods to fit such models, to test their validity, to estimate parameters, to predict future outcomes, etc.

A basic approach to this might be an augmented regression (linear) model. With data for 36 locations on a 6×6 rectangular grid measured monthly over several years, a model for log ozone level Z might take the form

$$z_{ijk} = \mu_{ijk} + \alpha_i + \gamma_{jk} + \varepsilon_{ijk}$$

for location (j, k) at time i (for a fixed discrete set of times and a prespecified and fixed grid of locations). Here μ_{ijk} is a regression component, with added random components $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\varepsilon})$ which we would need to test for zero means, independence, stationarity, no cross-correlations etc. The $\boldsymbol{\alpha}$ (and $\boldsymbol{\varepsilon}$) might need to be assumed to have common time-series representations, etc. In turn, this may not suffice. Thus we are effectively 'patching' together the space (regression) and time (time-series) approaches to produce a hybrid spatial/temporal analysis, but many other approaches could be (and have been) used even for this single and relatively straightforward problem. Landau and Barnett (1996) used such methods for interpolating several meteorological variables; see the discussion in the next subsection and Examples 11.5.

1.3.3 'Keep it simple'

Environmental problems are usually highly complex. We understand certain aspects of them, but almost never the whole picture. Thus general scientific considerations may suggest how a system should respond to a specific stimulus; observed data may show how it *did* react to some actual conditions. Examples include changes in river stocks of Chinook salmon over time (Speed, 1993) and, on a larger scale, the growth characteristics of winter wheat, a problem examined statistically by Landau *et al.* (1998).

Models for winter wheat growth could easily involve tens, hundreds or even thousands of parameters, including those representing weather characteristics throughout the growth period as well as the many relevant plant physiological features. Such highly parameterized models are quite intractable – we would be unable to fit parameters or interpret results. Occam's razor says 'in looking for an explanation, start with the simplest prospect'; such *parsimony* is vital in the complex field of environmental relationships.

Consider the wheat problem in more detail. Complicated, elaborate *mechanistic models* have been proposed (AFRCWHEAT2 for the UK, CERES for the USA and SIRIUS for New Zealand). These are based on claimed scientific (plant-growth) knowledge and simulated day-to-day growth and climatological conditions (but no actual data from real life). They attempt to represent the physical environmental system by means of a deterministic model, based on differential equations, which usually does not incorporate probabilistic or variational components, although it may in some applications be subjected to a 'sensitivity analysis'. A review of such modelling considerations in the context of the wheat models is given by Landau *et al.* (1998)

In global warming, the vast global circulation models for predicting climate change play a similar role – again they are essentially mechanistic (deterministic) in form.

Do they work? Some recent results by Landau *et al.* (1998) cast doubt on this, for the winter wheat models at least. A major data-assembly exercise was carried out to compile a database of wheat yields for about 1000 sites over many years throughout the UK where wheat was grown under controlled and well-documented conditions. These data constituted the 'observed yields' which were to be compared, for validation purposes, with the corresponding yields which the wheat models would predict. Crucial inputs for the models were daily maximum and minimum temperatures, rainfall and radiation as well as growth characteristics. These climatological measures were available for all relevant days and for 212 meteorological stations. But these meteorological stations were not, of course, located at the sites at which the wheat yields were measured (the wheat trial sites are shown in Figure 1.4). So it was necessary to carry out a major interpolation exercise for the meteorological variables. This is described by Landau and Barnett (1996) and was highly successful (see Figure 1.5, which shows actual *versus* interpolated minimum temperatures, where the fit accounts for 94% of the variation in observed values).

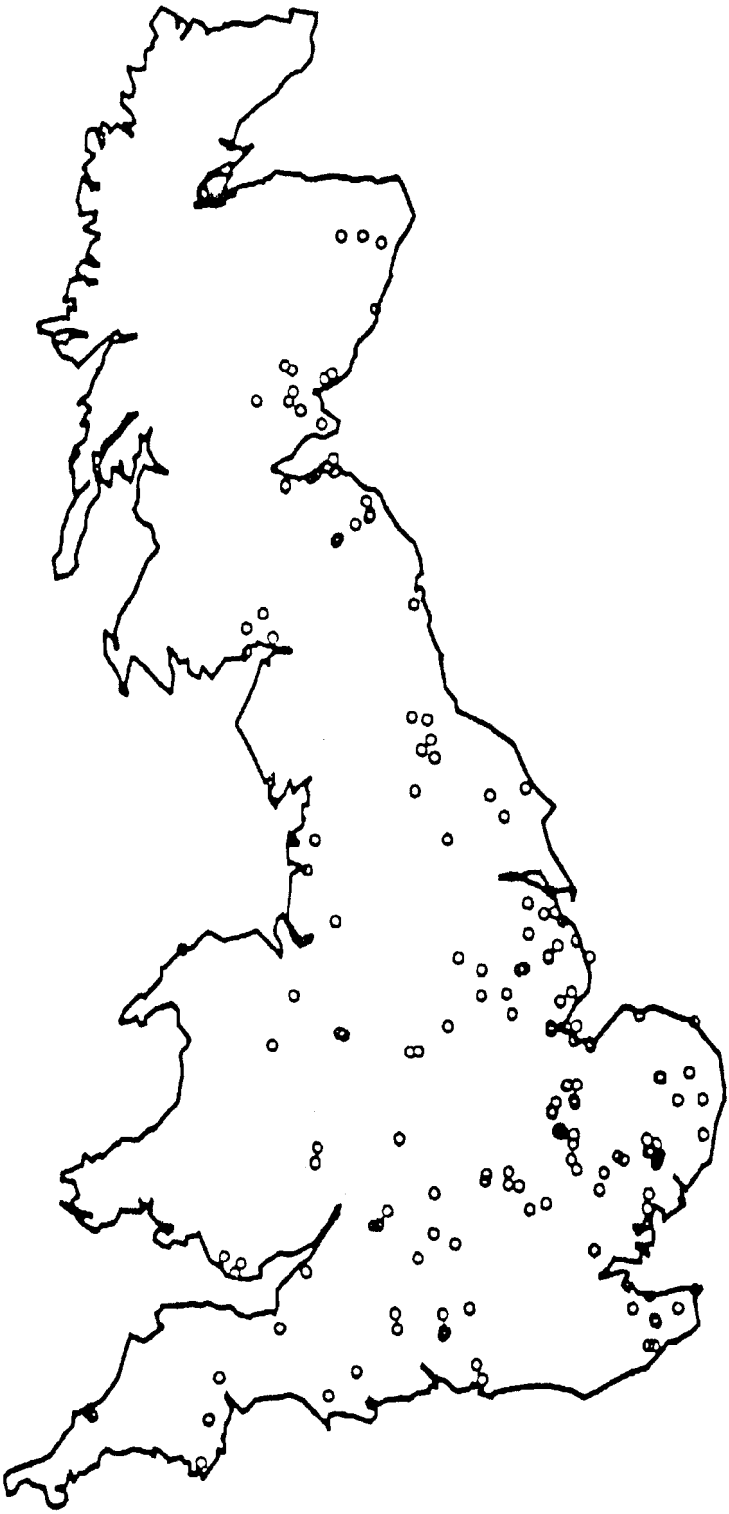

Figure 1.4 Sites of Wheat trials.

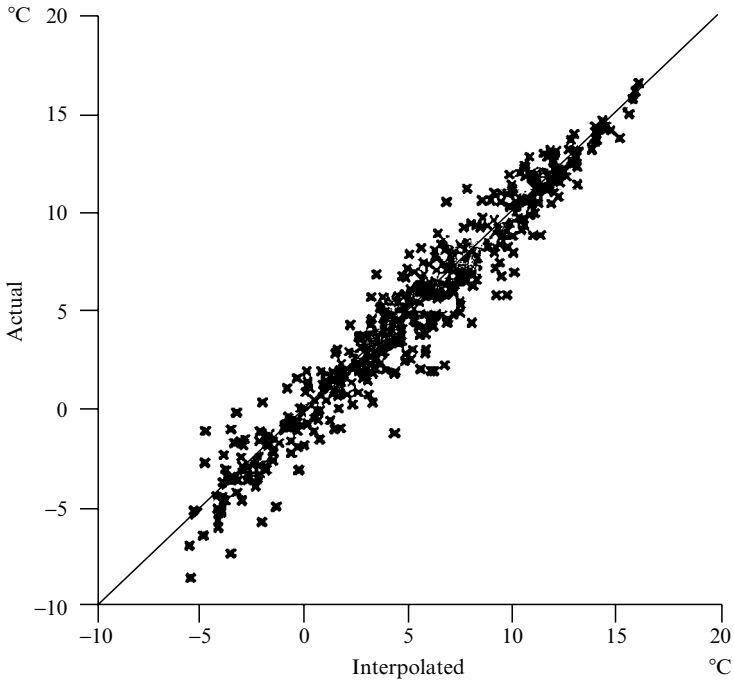

Figure 1.5 Actual and interpolated minimum temperatures.

When all culture and weather information was entered in the wheat models, they produced 'predicted yields' to compare with the 'observed' ones – the results were very surprising, as is shown by the plot (Figure 1.6, from Landau *et al*., 1998) of actual *versus* predicted yields for the AFRCWHEAT2 model. Such conspicuous lack of association arose with all the models!

At the opposite extreme of sophistication from mechanistic models (see discussion in Barnett *et al*., 1997), purely statistical (empirical) fits of regression-type models to real data may often predict environmental outcomes rather well but are sometimes criticized for not providing scientific understanding of the process under study. The ideal would presumably be to seek effective *parsimonious data-based mechanistic models*. Landau *et al*. (1998) went on to show how a parsimonious mechanistically motivated regression model predicted wheat yields with correlation in excess of 0.5.

1.3.4 'How much can we take?'

A range of environmental problems centre on the pattern of responses of individuals (or of systems) to different levels of stimulus – of patients to levels of a drug, of citizens to levels of pollution, of physical environment to 'levels' of climate (rain, wind, etc.). In such cases we might be interested in *trends*, in *dose–response relationships*, in *extremes*, and a variety of models and methods will be needed to address such questions as the maximum safe level for particulate

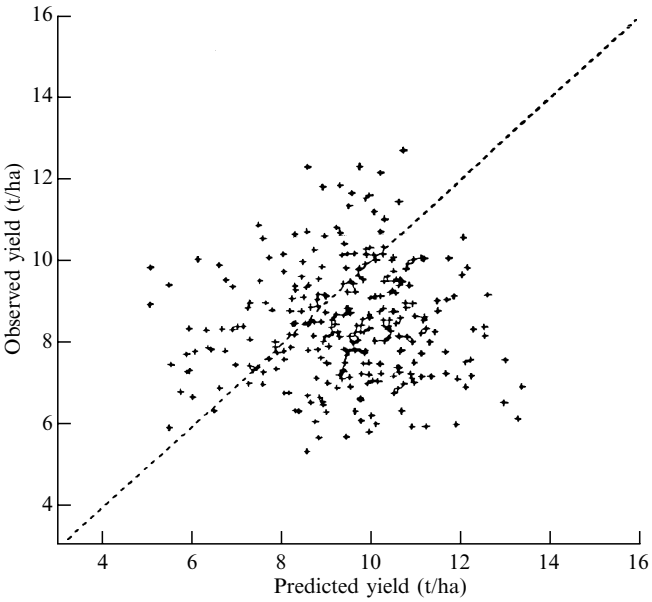

Figure 1.6 Observed versus predicted yields for AFRCWHEAT2 (Landau *et al.*, 1998 with permission from Elsevier).

matter from diesel fuel emissions, or how big a particular dam must be if it is not to overflow more than once in a hundred years?

Consider, as one specific example, a problem in the biomedical field where we are concerned with the control of harmful insects found on food crops by means of an insecticide, and wish to know what 'dose' needs to be used. One model for this is the so-called *logit* model (*probit* models are also used). The logit model expresses the proportion, $p(x)$, of insects killed in terms of the dose, or perhaps the log dose, x, of the applied 'treatment' in the form

$$\ln \frac{p(x)}{1-p(x)} = \alpha + \beta x.$$

In practice, we observe a binomial random variable $Y \sim \mathrm{B}(n(x), p(x))$ at each dose x which represents the number killed out of $n(x)$ observed at that dose level (with possibly different numbers of observations $n(x)$ at different doses). The general shape of this relationship – which is in the broad class of the *generalized linear model* (see Section 7.3) – is ogival (Figure 1.7). Of interest are such matters as the lethal effective doses needed to achieve 50% or 95% elimination of the insects; these are referred to as the LD_{50} and LD_{95}, respectively.

In a development of such interests another relevant topic is *bioassay*. This is concerned with evaluating the relative potency of two forms of stimulus by analysing the responses they produce in biological organisms at different doses. Typical data for what is called a *parallel-line assay* are shown in Table 1.1 (from Tsutakawa, 1982) in the context of examining the relative potencies of two treatment regimes applied at a range of dose levels to laboratory subjects. What

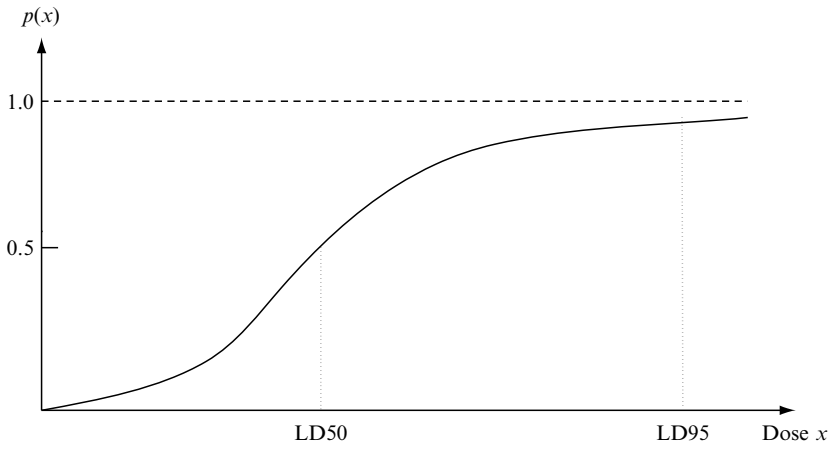

Figure 1.7 A dose–response curve.

would be of interest here is whether the two treatments show similar relationships with dose – apart, perhaps, from a scale factor. We pursue such matters further in Chapter 8.

Table 1.1 Bioassay data on rats: uterine weights (coded).

	Standard treatment			New treatment	
Dose	0.2	0.3	0.4	1.0	2.5
	73	77	118	79	101
	69	93	85	87	86
	71	116	105	71	105
	91	78	76	78	111
	80	87	101	92	102
	110	86		92	107
		101			102
		104			112

Source: Tsutakawa (1982).

1.3.5 'Over the top'

In the design of reservoirs to hold drinking water, it is important to contain and control the water by means of dams which are high and sound enough not to regularly overflow and thus cause loss of resource, and possible serious damage in the outflow area (Figure 1.8). In any year, the major threat of such damage occurs at the annual maximum water level X. If this exceeds the dam height h, we will have overflow and spillage on at least one occasion. It might seem sensible to design the dam to be of *great height*, h_0, with no realistic

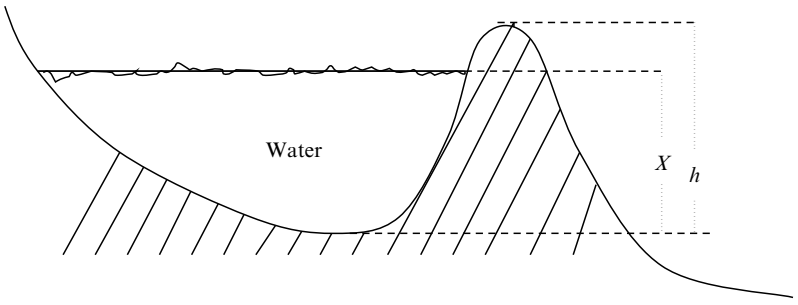

Figure 1.8 A reservoir and dam.

prospect that we will ever have $X > h_0$. But this would be unjustifiably expensive and we must seek a compromise dam height, h_1, where $P(X > h_1) = \gamma$ is designed to be acceptably small. A crucial feature of a dam's design is its *return period*, γ^{-1}. Thus if $\gamma = 0.01$, we would have a return period of 100 years.

Of course, X is a random variable describing the annual maximum water level. It will depend on a variety of features of the prevailing climate and the run-off terrain which drains into the reservoir. We do not know the distribution of X and at the design stage we may not even have any sample data from which to draw references about this distribution. But we do have some structural information about X from which to model it. It is an *annual maximum*. Thus, for example, if Y_i is a random variable describing the highest water level on day i, then $X = \max Y_i = Y_{(n)}$, with $n = 365$. Here, $Y_{(i)}$ is an *order statistic* – the ith largest of a set of identically distributed random variables $Y_i (i = 1, 2, \ldots, n)$. For the dam problem, $Y_{(n)}$ is the potential overall *yearly* maximum, and in a particular year it will take some specific value, $y_{(n)}$. Thus we have an example of a class of environmental problems where we are interested in *ordered random variables* and in particular in the largest, or *upper extreme,* of a set of more basic random variables. Such problems lead to an interest in the theory and methodology of *extremal processes*, a topic which we will examine in some detail in Section 2.3.

The reservoir problem exemplifies some of the features of this interesting area of study. We are unlikely to have much useful information about the characteristics of even our basic random variable Y: the assumed common random variable describing the *daily* maximum level. So what can we hope to say about the random variable of major interest, namely $Y_{(n)}$?

Interestingly, we can say something useful as a result of general *limit-law behaviour* of random variables which comes to our aid. If $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables then, *whatever their distribution,* it turns out that the distribution of the maximum $Y_{(n)}$ approaches, as $n \to \infty$, just one of three possible forms – the so-called *extreme value distributions*. One of these, the Gumbel extreme-value distribution, has a distribution function (d.f.) in the family

$$F_n(y) = \exp\{ -\exp[-(y - \alpha)/\beta]\}.$$

So, in spite of some departures from this structure (the Y_i may not necessarily be identically distributed, n is not infinite but is very large), it may be that we can model X for the dam problem by $F_n(x)$ above reducing our uncertainty to just the values of the two parameters α and β.

Consider another problem with similar features, where we are interested in a random variable Z describing the winter minimum temperature at some location. Here Z is a *minimum* rather than a maximum. So $Y = -Z$ is the *maximum* negative winter temperature, and we might again adopt a model which says that Y has d.f. $\exp\{-\exp[-(y-\alpha)/\beta]\}$.

So

$$P(Z > z) = P(-Y > z) = P(Y < -z) = \exp\{-\exp[(z+\alpha)/\beta]\}.$$

If we had observations $z_1, z_2, \ldots, z_n$, for n years, we could estimate $P(Z > z)$ by the empirical d.f. $(\#z_i > z)/(n+1)$. So if we order our observations as $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$ then we have the approximate relationship

$$\frac{n-i}{n+1} \approx \exp\left[\exp\left(\frac{z_{(i)}+\alpha}{\beta}\right)\right]$$

or

$$-\ln\left[\ln\left(\frac{n+1}{n-1}\right)\right] \approx \frac{z_{(i)}+\alpha}{\beta}.$$

So we could plot a graph of $z_{(i)}$ against $-\ln\ln[(n+1)/(n-i)]$ and expect to find a linear relationship. Such a probability plotting method was illustrated by Barnett and Lewis (1967) in the context of a problem concerned with the deterioration of diesel fuel at low winter temperatures. It was necessary to model the variation in Winter minimum temperatures from year to year at different locations. Figure 1.9 (from Barnett and Lewis, 1967) uses the above plotting technique to confirm the extreme-value distribution for three locations: Kew, Manchester Airport and Plymouth. The linearity of the plots provides compelling empirical evidence of the usefulness of the above extreme-value distribution.

1.4 FUNDAMENTALS

The notation and terminology used throughout the book will conform to the following pattern. A univariate random variable X will take a typical value x. The distribution of X will be represented in terms of a probability density function (p.d.f.) $f_\theta(x)$ (if X is continuous) or a probability function $p_\theta(x)$ (if X is discrete), or in either case by a distribution function $F_\theta(x)$, where θ is a scalar or vector parameter indexing the family of distributions in which that of X resides.Thus for discrete X, for example, we have

$$p_\theta(x) = P(X = x),$$
$$F_\theta(x) = P(X \leq x),$$

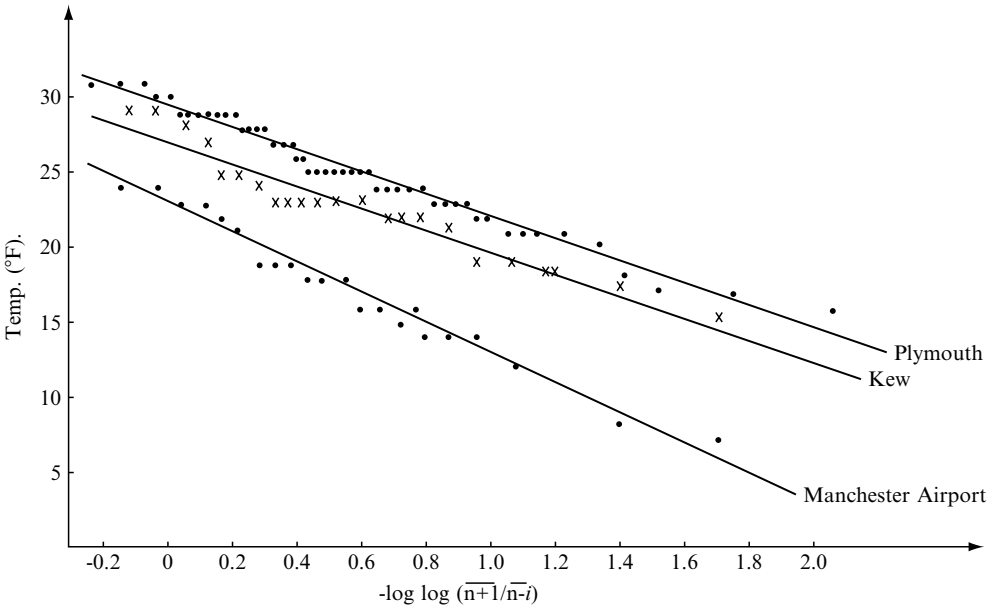

Figure 1.9 Low-temperature plots for typical locations (Barnett and Lewis, 1967).

where $P(\cdot)$ denotes probability. Common families of distribution of X include the discrete *binomial* and *Poisson* distributions and the continuous *normal, exponential* and *gamma* distributions. If X follows these respective distributions, this will be indicated in the following manner:

$$X \sim \mathrm{B}(n,p), \qquad X \sim \mathrm{P}(\mu)$$

and

$$X \sim \mathrm{N}(\mu,\sigma^2), \qquad X \sim \mathrm{Ex}(\lambda), \qquad X \sim \Gamma(r,\lambda),$$

where the arguments $(n,p), \mu, p, (\mu,\sigma^2), \lambda$ and (r,λ) are the symbols used in the distinct cases for the generic family parameter θ.

A random sample $x_1, x_2, \ldots, x_n$ of observations of X has mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is an unbiased estimator of the population parameter $\mu = \mathrm{E(X)}$ where $\mathrm{E}(\cdot)$ is the expectation operator. The sample variance,

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2,$$

is unbiased for the population variance $\sigma^2 = \mathrm{E}[(X-\mu)^2] = \mathrm{Var}(X)$. If we order sample observations as $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, the $x_{(i)}$ are observations of the *order statistics* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, We refer to $x_{(1)}$ and $x_{(n)}$ as the sample

minimum and sample *maximum* (or lower and upper *extremes*); $x_{(n)} - x_{(1)}$ is the *sample range* and $\left(x_{(1)} + x_{(n)}\right)/2$ is the *mid-range*. We will denote the sample *median* by m.

A multivariate random variable $\mathbf{X}$ of dimension p has components $X_1, X_2, \ldots, X_p$; a random sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is an $n \times p$ array

$$\begin{matrix} x_{11} & x_{21} & \ldots & x_{p1} \\ x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{pn} \end{matrix}$$

The distribution of $\mathbf{X}$ will be described, as appropriate, in terms of joint probability, or probability density, functions $p_\theta(\mathbf{x})$, or $F_\theta(\mathbf{x})$, or by its d.f. $F_\theta(\mathbf{x})$. For a bivariate random variable $\mathbf{X}$ it is often more convenient to represent its components as (X, Y).

It will be assumed that the reader is familiar with basic concepts and methods of probability theory and statistical inference, including correlation, regression, estimation, hypothesis testing, the maximum likelihood method, the maximum likelihood ratio test, etc., at a level covered by standard intermediate texts such as Mood *et al.* (1974) or Garthwaite *et al.* (1995).

1.5 BIBLIOGRAPHY

Throughout our discussion of the statistical models and methods and of practical examples of their use, in the different areas of environmental interest, we will need to consider important recent contributions in the literature. At all stages, references will be given to published papers in professional journals and to books or occasional publications. These will always be of specific relevance to the statistical topic or environmental application being studied or will enable it to be considered in a broader methodological or applications review setting. However, it is useful even at this early stage to refer to a small group of publications of general applicability and relevance.

Firstly, there are a few books which purport to cover the general theme of this book. These include elementary texts or sets of examples whose titles or descriptions indicate coverage of environmental statistics. These include Hewitt (1992), Berthouex and Brown (1994), Cothern and Ross (1994), Ott (1995), Pentecost (1999), Millard and Neerchal (2000) and Townend (2002). They are of varying level and range but do not usually go beyond basic statistical concepts and methods, with corresponding constraints on the types of applications and examples that can be discussed. A set of case studies on environmental statistics are presented in the edited volume by Nychka *et al.* (1998).

As indicated by the breakdown of topics into parts I–V explained in Section 1.2, we will be concerned with distinct areas of statistical principle and method. More detailed treatments are available of many of the areas and include the following:

Part I Galambos (1987) on extremes; David (1981) on order statistics; Barnett and Lewis (1994) on outliers; and Huber (1981) on robustness.

Part II Barnett (2002a) on survey sampling; Thompson (2002) and Wheater and Cook (2000) on general sampling methodology; and Thompson and Seber (1996) on adaptive sampling.

Part III Stapleton (1995) on linear models and regression; and McCullagh and Nelder (1989) on generalized linear models.

Part IV Barnett and O'Hagan (1997) on standards and regulations.

Part V Bloomfield (2000) on Fourier analysis of time series; Fuller (1995) on time series in general; Cressie (1993) on spatial models and methods; and Upton and Fingleton (1985, 1989) and Webster and Oliver (2001) on spatial data analysis.

Review papers can provide informative resumes or encapsulations of our theme: see, for example Barnett (1997). The *Encyclopaedia of Statistical Sciences* (Kotz *et al*., 1982–1988) provides crisp and informative explanations of many of the statistical topics discussed below, whilst the recent *Encyclopaedia of Environmentrics* (El-Shaarawi and Piegorsch, 2002) is to be welcomed for its comprehensive coverage of so much of the field of environmental statistics.