Chapter 15

STATISTICAL METHODS FOR ANALYSIS OF ENVIRONMENTAL DATA

G. Das

INTRODUCTION

Environmental Statistics is becoming an important discipline both for reasons of societal challenge and statistical opportunity. It is demanding more and more of non-traditional statistical approaches. This is partly because environmental studies involve space, time and innovative environmental sampling and monitoring. Also environmental statistics must satisfy environmental policy research in addition to disciplinary and interdisciplinary environmental research (Patil, 1994).

The aim of statistics as originally and still is to describe or characterise a population of individuals, objects or events (called units) based on certain well-defined measurements (called data) made on them. This lead to the development of descriptive data analysis (DDA), or methods of summarizing data through averages, measures of variation and association, and graphical representations. However, if we have observations only on a subset of units (called a sample), there

is some uncertainty in attributing the sample results to the population. How can the uncertain knowledge from a sample be used in taking decisions in real life? This question has baffled the human mind over a long period of time, and the breakthrough came with the realization that knowledge, however uncertain it may be, could be used in an optimum way (to minimize loss in decision making) if we know the amount of uncertainty in it. The main thrust of statistics shifted to what is called inferential data analysis (IDA) devoted to quantification of uncertainty in the results of a sample, based on specified stochastic models characterizing the differences between different possible samples. The paradigm of IDA is the logical equation (Rao, 1994):

Uncertain Knowledge	+	Knowledge of the amount of uncertainity	=	Usable knowledge

However the question arose: what is the uncertainty in the choice of stochastic model itself for IDA? What adjustment is needed in a specified model for possible defects in data such as measurement and recording error, wrong identification of units selected for measurement and any other unusual features of a sample? These considerations led to the concepts of robustness in the choice of statistical methodology to ensure validity of inference under a wide class of underlying stochastic models, and exploratory data analysis (EDA), or cross examination of data (CED) to use a more general term, to understand the nature of the data and detect possible defects for which adjustments may be necessary in the analysis. The various logical and methodological steps in modern statistical analysis from efficient data collection to comprehensive analysis and interpretation of results are exhibited in Figure 1.

We shall examine some specific issues, which arise in data analysis and discuss some techniques of statistical analysis for data related to environmental research.

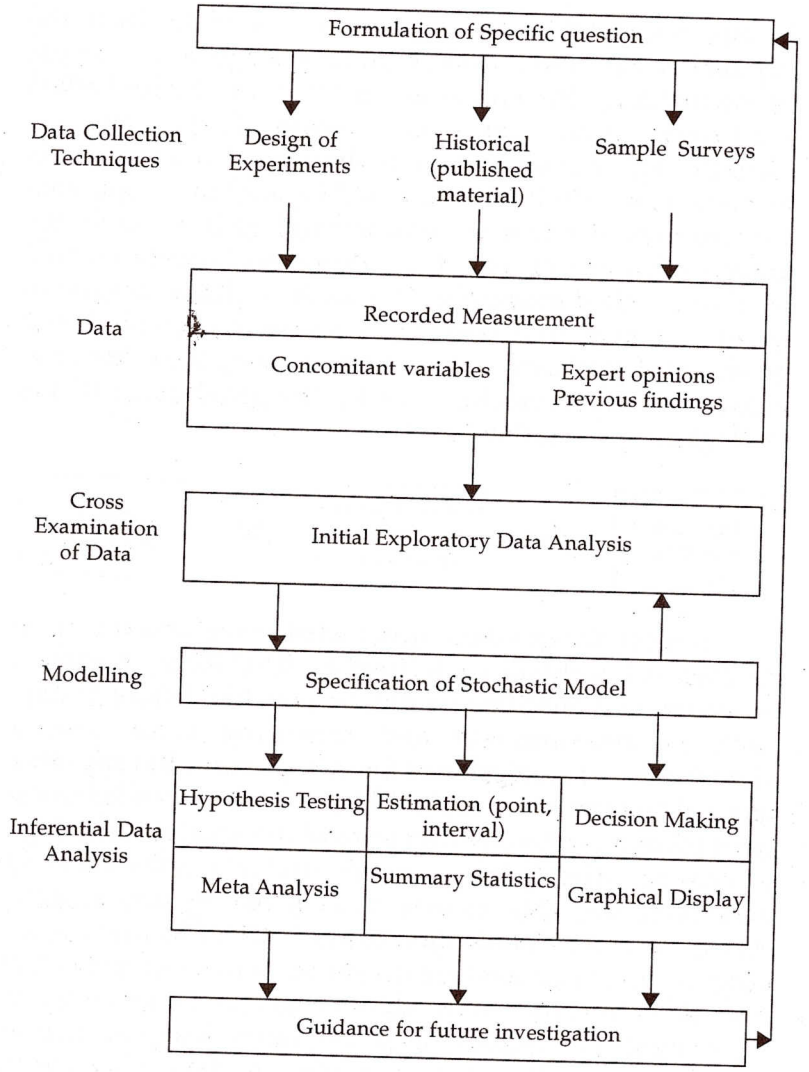

Figure 1: Statistical data analysis

ENVIRONMENTAL SAMPLING

Sampling consists of selection, acquisition and quantification of a part of the population. While selection and acquisition apply to physical sampling units of the population, quantification pertains only to the variable of interest. A sampling procedure is expected to provide a representative and informative sample. Considering this as desirable criteria a desired sample size is n or more. On the other hand, considerations of resources in terms of cost, time and effort usually lead to an affordable sample size m or less. This is known as 'cost-loss' dilemma. . A common experience is that m is much less than n. Thus what is desirable is not affordable and what is affordable is not adequate. Besides this there can be certain statistical considerations that necessitate a departure from the conventional methods of sample survey. Some of the important environmental sampling methods are:·

1. Encounter Sampling
2. Adaptive Sampling
3. Ranked set Sampling
4. Composite Sampling

A detail information about these methods may be obtained from Gore (1995).

EXPLORATORY DATA ANALYSIS (EDA)

Once the data is collected task of a statistician is to cross-examine the data before applying standard statistical techniques designed to answer specific questions. It pays to spend more time in initial analysis of data to understand its special features and to decide on an appropriate model for drawing inference. There is no general prescription for the initial data analysis, but to begin with following tabulations would be useful.

- Classify the samples into homogeneous sets using concomitant information if available and do the following.

Draw histogram or box plots or empirical distribution functions indicating the mean, median, the quartiles, the two highest and the lowest values, compute the first four moments and b_1 and b_2 the measures of skewness and kurtosis for each variable, draw bivariate scatterplots for every pair of variables and compute measures of association. Such an analysis would provide adequate information on the shapes of the distributions involved and presence outliers and clusters in the data.

- One can try transformation of variables and apply the procedures indicated above to see whether simpler models can be used in the analysis of data.
- Probability plots of data like the Q-Q plots will be useful in determining the adequacy of a given stochastic model and also the presence of any gaps in data.
- It would also be useful to subject the multivariate data to standard cluster algorithm techniques to determine whether there are distinct clusters in the data and whether the clusters should be treated as separate populations in statistical analysis.

INFERENTIAL DATA ANALYSIS (IDA)

Inferential data analysis refers to the statistical methodology based on a chosen stochastic model for observed data for estimation of unknown parameters, testing specified hypothesis, prediction of future observations, making decisions etc. The choice of a model may depend on the specific information we are seeking for the data. It may not necessarily be the one, which explains the whole observed data, but one, which provides efficient answers, to specified questions. Some of the relevant issues in the choice of a model are as follows:

- It is more illuminating to analyze given data under different possible alternative stochastic models and examine the difference in the conclusions that emerge.
- A robust procedure independent of any stochastic model should only be a last resort as robust procedures are usually inefficient.

- If possible the performance of any chosen model should be tested by cross validation using the same data under analysis. This may provide a basis for choosing an appropriate model for answering specific questions.
- Different stochastic models may be needed to answer different questions from the same data.
- Inferential data analysis should be of a interactive type: new features of the data may emerge during the analysis under a specified model requiring a change in the analysis originally contemplated.

The inferential data analysis stands for the entire body of statistical methods for estimation of unknown parameters, testing of special hypotheses, prediction of future events and decision making based on a chosen stochastic model. Data analysis should not be confined to answering specific questions raised by a customer. Data often contain valuable information to throw light on some other aspects of the problem under study and to indicate new lines of research for further expansion of knowledge. The main principle of data analysis can be enunciated in the form of an equation:

Data analysis = Answering Specific Questions
+ Extracting Information for New Lines of Research

TESTS OF SIGNIFICANCE

Is a null hypothesis true or false? Often this is a wrong question to ask. Once the set of all alternative hypotheses is specified, a more logical question to ask is which hypothesis or subset of hypotheses is the most plausible for the observed data. Testing whether a particular hypothesis is true is of limited use and is not of much value in decision-making. It is generally recommended in situations where alternatives cannot be specified.

Some illustrations of Inferential data analysis based on a chosen stochastic model for observed data for estimation of unknown parameters, testing specified hypothesis, prediction of future observations, making decisions etc. are given in the following sections.

NORMAL DISTRIBUTION

A probability or frequency law describes the stochastic behaviour of a variable quantity X. This law involves parameters such as mean (median or mode), variance, skewness among others. Once the probability law is assumed to follow a functional form, the object is to estimate the parameters.

One of the most commonly occurring laws is the well-known normal distribution with the functional form.

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}, -\infty < x < \infty$$

Here μ denotes the mean and σ^2 is the variance. The distribution has mean = median = mode and the distribution is perfectly symmetric about μ. The smaller the value of σ^2, more concentrated is the distribution about μ and tails approach zero very fast. In fact the probability mass beyond $(\mu - 3\sigma^2, \mu + 3\sigma)$ is less than 1%. For the data $(x_1, x_2, \ldots\ldots, x_n)$ a good estimate of μ is $\bar{x}$, the sample mean and of σ^2 is

$$\sigma^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$$

The normal distribution arises in a very natural way in case of repeated measurement model with

$$X_i = \mu + \varepsilon_i, i = 1,2\ldots\ldots\ldots, n$$

where ε_i is the error measurement which arises out of various factors beyond the control of the investigator. Some common examples are distribution of height, weight, pH levels in soil etc. The normal distribution arises as an approximation to some of the other models such as Gamma, Poisson, and Binomial among others. Normal distribution can also be obtained by considering a transformed variable Y = log X when X is skewed.

Suppose a variable X is known to be normally distributed in a given population with a known variance σ^2 but with an

unknown mean μ. Also, suppose it is suggested that the mean may be equal to a specified value, say μ_0, and we want to see how acceptable this suggestion is. We have then the hypothesis

$$H_0 : \mu = \mu_0$$

which needs to be verified. Such a hypothesis is called null hypothesis, because it states that there is no difference between μ and μ_0. The verification (or test) of H_0 has to be done on the basis of a random sample from this population. Let $x_1, x_2, \ldots\ldots, x_n$ be the values of X for random sample of size n, the observations being independent.

In order to test, let us assume, to begin with, that it is true — or that, in fact, the population mean of X is μ_0. From this assumption a number of results will follow. The most important result for our purpose is that $\bar{x}$ is, according to the assumption, normally distributed with mean μ_0 and variance σ^2/n – in other words $\sqrt{n}(\bar{x} - \mu_0)/\sigma$ is a normal deviate, Z. As such,

$$P\left\{\sqrt{n}|\bar{x} - \mu_0|/\sigma > 2.576\right\} = 0.01$$

To put it in a different way, in repeated sampling from this population, in only one in hundred samples is the value of $\sqrt{n}(\bar{x} - \mu_0)/\sigma$ expected to exceed 2.576 numerically. This fact then provides a test for the hypothesis. If in a given sample $\sqrt{n}|\bar{x} - \mu_0|/\sigma$ exceeds 2.576, then it means that a value has been obtained which is very improbable under the hypothesis. In such a case hypothesis itself will be held in suspicion. We say H_0 is *rejected*. On the other hand, if in the given sample $\sqrt{n}|\bar{x} - \mu_0|/\sigma$ does not exceed 2.576, i.e. if it takes a value which is not improbable under the hypothesis, one would find no reason to suspect the hypothesis. It would then be said to be *accepted*.

By acceptance of a hypothesis we do not mean it is proved to be true. All that is implied is that, so far as the given sample

is concerned, we find no reason to question the validity of the hypothesis. Nor does rejection of H_0 mean a disproof of H_0. It means simply that, in the light of given sample, H_0 does not seem to be a plausible hypothesis.

The mode of argument may be restated as follows: some difference between the sample mean $\bar{x}$ and the hypothetical population mean μ_0 is to be expected because of the inevitable sampling fluctuations. However if this difference were too large, say greater than $2.576 \times \sigma / \sqrt{n}$ –in other words, $\sqrt{n}|\bar{x} - \mu_0|/\sigma > 2.576$ — then one would say that it may not be due sampling fluctuations alone but arises because the true population mean is not μ_0. One would thus take it as significant or indicative of the falsity of the hypothesis·

A test of this kind is called a *test of significance*. The probability 0.01, on the basis of which the differences are being regarded as significant or not, is called the *level of significance.* The choice of the level of significance, of course, depends on the experimenter himself. If he thinks that rejection of the hypothesis when actually it is true will be a serious error, he will chose a rather small value, say 0.01or 0.001. On the other hand, if he thinks this error is not so serious, he will not mind taking a high value as high as, say. 0.05 or 0.1. The general symbol for the level of significance is α (Goon, 1983).

Confidence intervals for level I –α for μ, the mean are given by

$$\left(\bar{x} - Z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

where Z_p denotes 100p % point of the standard normal distribution. For conventional levels $\alpha = 0.05$ and $\alpha = 0.01$, the values of $Z_{\alpha/2}$ are 1.96 and 2.576 respectively. Note that the interpretation of the confidence interval is that $\bar{x} \pm Z_{\alpha/2}\frac{s}{\sqrt{n}}$ will cover the true value of μ in 100(1–α) % cases.

Example 1. (Kunte, 1995) In 42 measurements on "Trace metals in sea scallops" the amount of cadmium in sea scallops

at number of different stations gave $\bar{x} = 9.28$ and s = 4.68. Then 95% confidence interval for the population mean μ is

$$9.28 \pm 1.96 \times \frac{4.68}{\sqrt{42}} = 9.28 \pm 1.42 = (7.86, 10.7)$$

BINOMIAL DISTRIBUTION

We consider a series of trials/experiments and observe the occurrence of an event (E) or (not E). If the trials are independent i.e. occurrences of E or not E in the previous trial does not effect the outcome of the next trial then probability of observing r successes (i.e. occurrences of E) in n trials is given by

$$B(r, n, p) = \frac{n!}{r!(n-r)!}p^r q^{n-r}, r = 1,2,\ldots\ldots\ldots n,$$

$$o < p < 1, q = 1 - p$$

original interest in the model arose out of concern with the problem as to whether male or female births occur equally frequently. A very reasonable estimate of p is

$$\hat{p} = \frac{r}{n} \text{ with } E(\hat{p}) = p \text{ and } Var(\hat{p}) = \frac{p(1-p)}{n}$$

Note that if

$X_i = 1$, if E occurs at ith trial and
$X_i = 0$, otherwise

then $T = \Sigma X_i$ = frequency of E in n trials and $\hat{p} = \bar{x}$ and var $(\hat{p}) = \frac{p(1-p)}{n}$ is estimated by

$$\frac{\hat{p}(1-\hat{p})}{n} = \frac{r(n-r)}{n^3}$$

Example 2. (Kunte, 1995) A very interesting use of Binomial model is in estimating size of wild life populations.

For example consider the problem of estimating the population size of fur seal pups. In one such rookery in Alaska in 1961, 4965 fur seal pups were caught and tagged in early August 1961. In late August, 900 fur seal pups were caught out of which 218 were tagged. Let N be the total number of fur seal pups in August 1961 in that area which we are trying to estimate. Then $\frac{4965}{N} = \hat{p}$ is a good estimate of p_1 the probability of fur seal being caught in that area in the first catch. Now in the second catch we have $\frac{218}{900}$ is a good estimate of p_2 the probability that a tagged pup is recaptured again. It is reasonable to assume that $p_1 = p_2$ and thus $\frac{218}{900}$ and $\frac{4965}{N}$ estimate the same parameter and therefore

$$N = \frac{4965 \times 900}{218} = 20497.706$$ or 20,500 to the nearest hundred.

Of course there are several assumptions made here.

1. The tagging does not affect the probability of catch and tags do not come off.
2. The total population of fur seal pups does not change significantly during the two catches either by births, deaths and migration.
3 The probability of catch is same all over the area.

POISSON DISTRIBUTION

Consider a small interval of time or space. Let E be the event that occurs in that interval with $P(E) = p$. If x is the number of occurrence of E in the total time or space then

$$\mathbf{P}(x,\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!}, x = 0,1,2.....$$

where number of trials n is very large, p is very small and $\lambda = np$, and mean = variance = λ.

This distribution plays an important role in detecting non-randomness in vegetation. When the individuals tend to be clumped together, the distribution of individuals is said to be *contagious*. And when the individuals are scattered very evenly over the area the distribution is said to be *regular*.

Example 3. (Kershaw, 1973) The following data shows the frequency distribution for number of individuals in each quadrat for 100 random samples taken from a community.

Number of individuals in each quadrat (x_i)	0	1	2	3
Frequency of occurrence in 100 quadrats (f_i)	46	34	14	6

For detecting the non-randomness in vegetation the observed number of individuals per quadrat are compared with the expected number derived from the Poisson distribution.

$$\mathbf{P}(x,\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!}, x = 0,1,2.....$$

Here λ is mean density of individuals. For this data the mean density

$$\lambda = \frac{\Sigma f_i x_i}{100} = 0.8$$

The expected number of quadrats containing 0, 1, 2, ...individuals are given in the following table. For testing the significance of difference between the observed and expected frequency chi-square test of goodness of fit was employed.

Number of individuals per quadrat	0	1	2	3
Observed frequency	46	34	14	6
Expected frequency	44.9	35.9	14.4	3.8
Difference	1.1	1.9	0.4	2.2

Chi-square is calculated as the differences squared divided by the expected frequency:

$$\chi^2 = \frac{(1.1)^2}{44.9} + \frac{(1.9)^2}{35.9} + \frac{(0.4)^2}{14.4} + \frac{(2.2)^2}{3.8} = 1.4123$$

The χ^2 table is entered at 2 degrees of freedom (2 less than the number of terms used to calculate χ^2 total) which shows a value of 1.386 when p = 0.5. Thus the chance of this difference between the two sets of frequencies arising accidentally is quite high. Therefore we can regard the observed data as showing a very good fit with the expected frequencies. Hence the population sampled was randomly distributed.

On the contrary consider a similar data from a different community as follows.

Example 3. (Kershaw, 1973) The following data shows the frequency distribution for number of individuals in each quadrat for 100 random samples taken from a community.

No. of individuals in each quadrat	0	1	2	3	4	5	6	≥7
Frequency of occurrence in 100 quadrats	47	6	5	8	5	6	7	16

The expected number of quadrats containing 0, 1, 2, ...individuals are obtained and then chi-square is calculated as the differences squared divided by the expected frequency:

$$\chi^2 = 470.334,\ p < 0.001$$

This shows that the chances of the difference between the observed and expected numbers arising accidentally is very less, and hence the difference is highly significant.

REGRESSION ANALYSIS

There is considerable literature on estimation of the regression function and its applications. There are robust procedures to take care of outliers, heteroscedasticity and multicollinearity. There are methods for the selection of predictor variables, determining influential observations, transformations of variables and so on.

However, there are other problems, which should be considered in establishing relationships between variables for prediction purposes or understanding the structure of the variables. Perhaps, the first step is to find out, using cluster analysis algorithm, whether there are gaps and distinct clusters in the data of predictor variables and in the combined data of criterion and predictor variables. If there are distinct clusters, the possibility of considering each cluster separately for analysis should be explored. There is a possibility of establishing a wrong association if we disregard the presence of clusters. There is also much information to be gained in explaining the existence of clusters and their inter-relationships.

The basic results that are used for making inferences in the context of regression model are based on summary statistics computed from the data. The results are valid and have meaning only in so far as the assumptions concerning the residual terms in the model are satisfied. A simple and effective method of detecting model deficiencies in regression analysis is by examining the residuals (Chatterjee & Price, 1977). The ith residual is defined as

$$e_i = y_i - \hat{y}_i$$

Corresponding to e_i, we have also defined the ith standardized residual

$$e_{is} = \frac{e_i}{s}$$

where s is the standard deviation of residuals. The standardized residuals have zero mean and unit standard deviation. With a moderately large sample, these residuals should be distributed approximately as independent normal deviates. An appropriate graph of the residuals will often expose gross model violations when they are present. Some of the more commonly used plots are those in which the standardized residuals e_{is} are plotted as ordinate against

- The fitted value
- The independent variable
- The time order in which the observations occur

as abscissa. In general, when the model is correct, then standardized residuals tend to fall between 2 and –2 and are randomly distributed about zero. The residual pattern should show no distinct pattern of variation. Besides there are some standard tests for examining the validity of the assumptions made in the model viz. (Prajneshu, 1997):

- One sample run test for randomness of residuals
- Shapiro-Wilk test for normality of residuals
- Durbin-Watson test for autocorrelation in residuals

To take a final decision about the appropriateness of a model, goodness of fit statistics R^2 , Root Mean Square Error (RMSE), Mean Absolute Error (MAE) etc. may be computed.

USE OF GRAPHICS

Graphical methods and displays have an important role in statistical data analysis. They enable the statistician to understand the nature of the observed data and decide on an appropriate statistical analysis. They are useful in drawing tentative conclusions, which can be verified by inferential data analysis. Graphs are also needed in communicating final results. There are numerous software programs for graphical representation of data and viewing high dimensional data in various projections into two and three-dimensional spaces. This is probably the most effective way of detecting gaps and clusters in data. Some innovations may have to be made in the existing programs to deal with particular data sets having some special features.

CONCLUDING REMARKS

Statistical methods were initially developed for use in biology, demography and behavioral sciences and later in business and quality control in industrial production. While the basic methodology of estimation, testing of hypothesis and decision making is common to all areas, there are special techniques developed to answer specific questions in each area. Environmental Statistics is a relatively new subject and as in

other new areas of application there will be a need to develop special statistical techniques for use in environmental studies. Strategies for sampling over time and space, spatial processes, search for hot spots, Bayesian methods, acceptance sampling and encounter sampling are some of the topics which need to be developed with special reference to problems in environmental monitoring and research.

REFERENCES

Chatterjee, Samprit and B. Price. 1977. *Regression Analysis by Example*, John Wiley & Sons.

Goon, A.M., M.K. Gupta and B. Dasgupta. 1983. *Fundamentals of Statistics*, Vol. I, World Press, Kolkata.

Gore, S.D. 1995. Environmental Sampling. In *Statistical Methods for Ecologists*, A.P.

Gore; M.B., A.V. Rajarshi, Kharshikar and S.A. Paranjpe (eds.), Regional Workshop, Deptt. of Statistics, Univ. of Poona.

Kershaw, Kenneth A. 1973. *Quantitative and Dynamic Plant Ecology*. Edward Arnold, London.

Kunte, S. 1995. Statistical Inference. In *Statistical Methods for Ecologists*, A.P. Gore, M.B. Rajarshi, A.V. Kharshikar and S.A. Paranjpe (eds.), Regional Workshop, Department of Statistics, Univ. of Poona.

Patil, G.P. 1994. Editorial, *Environmental and Ecological Statistics*, 1, 1–6.

Prajneshu. 1997. Von Bertalanffy growth model with autocorrelated errors, Indian J. Fish., **44(1)**:63–67.

Rao, C.R. 1994. Statistics: an essential technology in environmental research and management, *Environmental and Ecologycal Statistics*, 1, 7–19.