Chapter Six
Data Analysis and Interpretation

6.1. Introduction

This chapter provides guidance on the use of common statistical methods for the analysis of water quality data. The information is at an introductory level that will help the monitoring team identify suitable methods of analysis and assist with the interpretation of results. Much of the technical detail and the more advanced statistical procedures have been relegated to Appendix 5 where, in most cases, they are illustrated with the help of worked examples that demonstrate options available, methods of implementation and interpretation of results. The information provided in this chapter is not exhaustive; complex studies may require a greater level of statistical sophistication than is presented here, and for these studies the monitoring team is advised to consult a professional statistician.

Data analysis should be viewed as an integral component of the water quality management process. A framework for undertaking data analysis and interpretation is shown in Figure 6.1. A checklist is presented in Table 6.1.

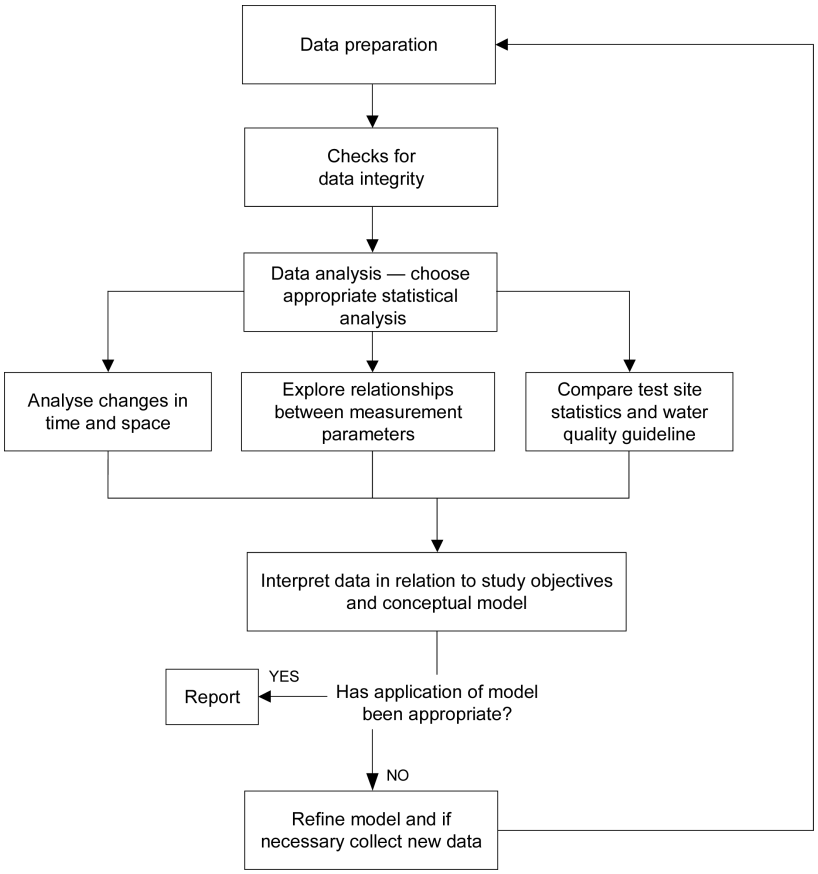

Figure 6.1. Framework for data analysis and interpretation

Table 6.1. Checklist for data analysis

1.	Preliminaries before data analysis; data preparation
	(a) Has the purpose of the data analysis exercise been clearly stated?
	(b) Have the parameters to be estimated and/or hypotheses to be tested been identified?
	(c) Will data from different sources be compatible (levels of measurement, spatial scale, time-scale)?
	(d) Have objectives concerning quality and quantity of data been articulated?
	(e) Has a program of statistical analyses been identified?
	(f) Have the assumptions that need to be met for correct application of statistical methods been listed?
	(g) Has data organisation (storage media, layout, treatment of inconsistencies, outliers, missing observations and below detection limit data) been considered?
2.	Have data reduction methods (graphical, numerical, and tabular summaries) been applied?
3.	Have 'aberrant' observations been identified and remedial action taken?
4.	Have potential violations of statistical assumptions (e.g. non-normality, non-constant variance, autocorrelation) been checked for?
5.	Have data been suitably transformed if necessary?
6.	Have data been analysed using previously identified methods; have alternative procedures been identified for data not amenable to particular techniques?
7.	Have results of analysis been collated into a concise (statistical) summary and have statistical diagnostics (e.g. residual checking) been used to confirm the utility of the approach?
8.	Has the statistical output been carefully assessed and given a non-technical interpretation?
9.	Have the objectives been addressed? If not, has the study been redesigned, have new or additional data been collected, the conceptual models refined and data re-analysed?

As noted in earlier chapters, data types, quantities, and methods of statistical analysis need to be considered collectively and at the early planning stages of any monitoring strategy. Decisions must be made about:

- measurement scales,
- frequency of data collection,
- level of replication, and
- spatial and temporal coverage,

so that data of sufficient quality and quantity are collected for subsequent statistical analysis. It is also important for the monitoring team to avoid the 'data rich–information poor' syndrome of collecting data that will not be subsequently analysed or that do not address the monitoring objectives.

Given the costs associated with the data collection process, it is imperative for the monitoring team to use formal quality assurance and quality control (QA/QC) procedures to ensure the integrity of the data. These procedures should be supported by established back-up and archival processes. There should be serious consideration of the archival medium to be used, because rapid advances in computer technology tend to increase the speed at which equipment becomes obsolete. The quality assurance of any monitoring program should include the data analysis, as well as the field and laboratory practices. Before statistically analysing the monitoring data, the monitoring team should use standard methods of data summary, presentation, and outlier checking to help identify 'aberrant' observations. If undetected, these data values can have profound effects on subsequent statistical analyses and can lead to incorrect conclusions and flawed decision making.

The monitoring team needs to develop a plan of the sequence of actions they will use for the statistical analysis of the water quality data. Only some of the many available statistical techniques need be used. An initial focus of monitoring might be to assess water quality against a guideline value or to detect trends, but an ultimate objective of the data analysis exercise will probably be to increase

the team's understanding of the natural system under investigation. Improved understanding should result in more informed decision making which in turn will lead to better environmental management. One of the most significant challenges for the data analysis phase is to extract a 'signal' from an inherently noisy environment.

As discussed in section 3.2 (Chapter 3), monitoring study designs will fall into three basic categories: descriptive studies including audit monitoring; studies for the measurement of change, including assessment of water quality against a guideline value (although that sort of study can also be categorised as descriptive); and studies for system understanding, including cause-and-effect studies and general investigations of environmental processes. The statistical requirements for the descriptive studies are less complex than for the other categories in which more detailed inferences are being sought.

Most of the statistical methods presented in this chapter are based on classical tools of statistical inference (e.g. analysis of variance, *t*-tests, *F* tests, etc.). These methods have served researchers from a variety of backgrounds and disciplines extremely well over many years. However, there is growing concern about their utility for environmental sampling and assessment. When natural ecosystems or processes are being measured it is invariably hard to justify the assumptions that response variables are normally distributed, that variance is constant in space and time, and that observations are uncorrelated. In these cases, remedial action (e.g. data transformations) may overcome some of the difficulties, but it is more probable that an alternative statistical approach is needed. For example, generalised linear models are often more suitable than classical ANOVA techniques for the analysis of count data because of their inherent recognition and treatment of a non-normal response variable. Recently, a number of researchers have questioned the utility and appropriateness of statistical significance testing for environmental assessment (e.g. McBride et al. 1993; Johnson 1999; and see section 6.4.2). Also, there is a progressive move towards more 'risk-based' methods of water quality assessment; although some of these methods also have difficulties (Fox 1999).

A large number of introductory and advanced texts of statistical methods are available (e.g. Ott 1984; Helsel and Hirsch 2000). For a relatively detailed and comprehensive description of statistical techniques used in water quality management studies, see Helsel and Hirsch (2000).

Statistical formulae are presented in this chapter and in Appendix 5 for clarification or for completeness, but it is recognised that the monitoring team will consign most computation to a statistical software package.

There is a plethora of statistical software tools on the market and it is beyond the scope of the Monitoring Guidelines to review them all. However, several are worthy of mention. SAS® is a powerful data analysis tool suited to handling large data sets; MINITAB®, STATISTICA® and SYSTAT® are suited to the analysis of medium–large data sets and their popularity stems from ease of use and a comprehensive set of procedures; S-PLUS® embodies many contemporary and robust statistical procedures which are not readily available elsewhere. Microsoft EXCEL® is useful for data screening and manipulation. EXCEL® has a limited number of intrinsic statistical functions, and a number of third-party statistical 'add-ins' are also available. Concerns have been raised about the accuracy and numerical stability of the statistical algorithms in EXCEL®; based on the results of rigorous testing, McCullough and Wilson (1999) concluded that 'persons desiring to conduct statistical analyses of data are advised not to use EXCEL'. However, while this is important when accuracy to many decimal places is needed or when algorithms must deal with 'ill-conditioned' data sets, it is unlikely to have measurable effects on the types of analyses contemplated by water quality scientists and natural resource managers.

All these software tools provide a high level of functionality and technical sophistication, but they also lend themselves to abuse through blind application. The monitoring team should not indulge in unstructured and undisciplined application of techniques because they 'seem to work' or produce a predetermined outcome. However, the team may find Exploratory Data Analysis (EDA) and data mining useful.

The remainder of this chapter outlines the data analyses ordinarily undertaken in practice. The first part recommends and suggests techniques for summarising data sets and reducing them to descriptive numerical quantities, and for data visualisation, transformations, outlier detection, censoring, trend detection and smoothing. Statistical techniques referred to here are designed to help tease out patterns in data sets, identify anomalies, and condense raw data using graphical tools and summary statistics. Later sections of the chapter cover a wide range of 'classical' statistical methods for helping the monitoring team make decisions about the likely values of water quality parameters, observe changes in these parameters over space and time, and explore relationships between pairs or groups of parameters. The later sections also refer to contemporary statistical techniques such as generalised additive models, robust regression and smoothing; these are usually computationally intensive and can only be realistically undertaken with a fast computer and appropriate statistical software.

6.2. Data Preparation

The data obtained from laboratory and field studies need to be summarised in a form that is amenable to analysis (see also section A5.1.1). It is best to check that the data provided are only those that are acceptable by field and laboratory quality assurance/quality control (QA/QC) criteria, and that they are rounded off to the appropriate number of significant figures.

Analytical data can be tabulated into spreadsheets that are immediately useable for exploratory analysis. Physical measurements should be tabulated in a form that permits ready comparisons with chemical and biological data for the same sites. The choice of formats for these is reasonably intuitive.

Graphical presentations of the basic data are also useful for displaying differences; e.g. a profile can illustrate changes in metal concentrations with depth in a sediment core.

Prior to more comprehensive analysis, the data must be given a preliminary examination to check their integrity before they can be subjected to more detailed analysis. Data that are missing or below detection limits (so-called 'censored' data) will need to be considered, as will obvious outliers that might be attributable to experimental or transcription errors.

6.2.1. Censored Data

Unless the water body is degraded, failure to detect contaminants is common. Rather than concluding that the particular contaminant does not exist in a water sample, the monitoring team records the observation as 'below detection limit' (BDL). Unfortunately, there is no universally accepted method of dealing with BDL data. Some common approaches include these:

- treat the observation as 'missing';
- treat the observation as zero;
- use the numerical value of the detection limit;
- use the numerical value of half the detection limit.

When a large portion of the data is below detection limit, use of any of the above approaches will be problematic because the sample variance will be severely underestimated. Also, when standard statistical techniques are applied to data sets that have constant values in place of the BDL values, the resulting estimates are biased (El-Shaarawi and Esterby 1992).

Assigning a missing value code (such as * or NA) can also cause difficulties because the various software tools treat missing values differently. For example, some software packages compute an average using the full data set with missing values replaced by zeros (not a good strategy) while others simply ignore all missing values and reduce the sample size accordingly. More sophisticated statistical procedures have been devised that use re-sampling or imputation techniques to 'infer' a reasonable value for the BDL observation; their implementation requires advanced statistical skills.

Gilliom and Helsel (1986) and Helsel and Gilliom (1986) have estimated water quality parameters with censored data using a variety of techniques, while Liu et al. (1997) describe a method based on 'forward censored regression' to model data with censored observations. Commercial software packages for 'data imputation' have recently become available, although their narrow focus and restricted audience will probably not have any measurable impact on the way environmental managers treat BDL data.

In the absence of more sophisticated tools for analysing censored data it is suggested that routine water quality parameters (means, percentiles, etc.) be computed using the full data set with BDL data replaced by either the detection limit or half the detection limit. The impact of this strategy on computed statistical measures should be clearly understood, and the monitoring team should *not* proceed with any form of inferential analysis (e.g. confidence intervals or hypothesis testing) when a significant proportion (e.g. >25%) of the data set is BDL and has been substituted by a surrogate value. Advanced statistical skills should be sought when any form of inference is required in these situations.

If only a small proportion of the data set is BDL and has been replaced by a numerical surrogate, it is best to perform any statistical analysis twice, once using zero and once using the detection limit (or half the detection limit) as the replacement value. If results from the two analyses differ markedly, the monitoring team should investigate more sophisticated statistical methods of dealing with censored observations (e.g. software packages LIMDEP by Econometric Software and SHAZAM by University of British Columbia). If the results do not differ markedly, the censored observations probably have little influence on the analysis.

6.2.2. Data Integrity

The integrity of water quality data can be reduced in many and varied ways. Losses or errors can occur at the time of collection, and in the laboratory during sample preparation and analysis, and during recording of results, and during electronic manipulation and processing, and during analysis, interpretation and reporting. Once the 'certified' data leave the laboratory, there is ample opportunity for 'contamination' of results to occur. Gross errors that are probably the result of data manipulations (transcribing, transposing rows and columns, editing, recoding, and conversion of units) are easily overlooked unless a modest level of screening is undertaken. While these sorts of errors can usually be detected by a scan of the raw data, more subtle effects (e.g. repeated data, accidental deletion of one or two observations, or mixed scales) are more difficult to identify. If left unchecked, these 'anomalous' readings can have a profound impact on subsequent statistical analyses and possibly lead to erroneous conclusions being drawn.

A good QA/QC program for data analysis uses simple yet effective statistical tools for screening data as they are received from laboratories. These typically include a mixture of graphical procedures (histograms, box-plots, time sequence plots, control charts, etc.) and descriptive numerical measures of key distributional aspects (mean, standard deviation, coefficient of variation, and possibly measures of skewness and kurtosis — see Appendix 5 (e.g. section A5.1.1) for a more complete description of these). These analyses are routine, but should not be ignored. Neither should they be unsupervised. The data analyst should oversee all processing of the data and carefully inspect graphs and reports. Most if not all pre-processing, manipulation and screening can be undertaken using the built-in capabilities of an electronic database system. The treatment of 'unusual' observations or 'outliers' is more contentious and is discussed next.

Care should be exercised in labelling extreme observations as 'outliers'. An outlier is indeed an extreme observation, although the converse is not necessarily true. It should be kept in mind that about two out of every 1000 observations from a normal distribution will fall beyond three standard deviations from the mean. To automatically label these as outliers and discard them from the data set would introduce bias into subsequent analyses. On the balance of probabilities, an observation beyond three standard deviations from the mean is likely to be 'aberrant'. Such observations need to be

highlighted for follow-up investigation to identify causes (e.g. recording error, laboratory error, abnormal physical conditions during sampling).

If there is no rational explanation for its value, the decision to include or exclude an outlier from the data set rests with the data analyst. It is suggested that only the most extreme observations (e.g. those that are four or more standard deviations from the mean) be excluded unless other good reasons can be established (e.g. different analytical method used). There are statistical tests for determining if a specific value can be treated as an outlier (see Neter et al. 1996) and techniques such as the box-plot can help in the decision. However outliers may convey significant information and their presence should initiate more thorough investigation[1]. Simple descriptive statistical measures and graphical techniques, combined with the monitoring team's knowledge of the system under investigation, are very valuable tools for identifying outliers.

Identification of aberrant observations in a multivariate (i.e. many variable) context is more complex. Water quality monitoring generates measurements of, say, metals, nutrients, organic compounds and other compounds, so for each sample there is often a vector of observations rather than a single value. These values (variables) tend to be correlated among themselves to some extent, indicating that they co-vary. If the co-dependence between variables is ignored and the aberrant observations are examined one variable at a time, unusual observations may be missed, and the whole procedure will be inefficient. In a multivariate context, it is quite possible for an observation to be 'unusual' even when it is reasonably close to the respective means of each of the constituent variables.

It is even more difficult to determine the causes of outliers in a multivariate context. Garner et al. (1991) have considered statistical procedures to help in this evaluation, but again care needs to be exercised in the labelling and treatment of potential outlying observations. The issue of outliers and multivariate normality for compositional data (i.e. where data represent percentages of various chemical constituents) has been investigated by Barceló et al. (1996). See also section A5.1.3.

6.3. Exploratory Data Analysis (EDA)

The way in which data are analysed will largely depend on the type of study being undertaken. The framework in Figure 6.1 suggests paths for analysing data from descriptive studies or from studies that measure change or system understanding; these will largely have been decided in the study design process, which defined the particular questions that need to be addressed (Chapter 3). This section discusses the range of data processing tools that might now be used.

6.3.1. Data Reduction

Data reduction is an important first step in the data analysis process. It enables the monitoring team to present and summarise important features of the data, and it also helps the analyst identify outliers or observations that are 'aberrant' in some other way.

A combination of statistical tools can be used for data reduction, including graphs (e.g. histograms, box-plots, dot-plots, scatterplots), tables (e.g. frequency distributions, cross-tabulations), and numerical measures (e.g. means, medians, standard deviations, percentiles). The objective of calculating summary statistics is to convey the essential information contained in a data set as concisely and as clearly as possible and/or to estimate a parameter of some population of values (see also section A5.1.1).

Frequency tabulations are commonly used for summarising data because they can condense even large data sets to manageable form without substantial loss of information, and because they can be graphically presented in the form of histograms and bar-charts.

[1] There is a salutary lesson to be learned from the automatic discarding of outliers. The hole in the ozone layer went initially undetected because computers used in the processing of the data had been programmed to remove extreme observations prior to analysis.

It is common to use statistics that are measures of the central tendency or 'average', such as the arithmetic mean, the median and the mode. A description of some common 'averages' is given in Table 6.2. For most water quality applications the arithmetic mean, geometric mean or median are the most appropriate quantities to use. The median is a robust estimator of central tendency because it is relatively unaffected by extremes in the data. The arithmetic mean does not share this property, but it is nevertheless the most widely used 'average': it is easy to compute and uses all the data. Also, it has well-established statistical properties, and many statistical tests relate to inference about a population mean. However, the choice of measure should not be automatic and will depend on the circumstances at hand. The arithmetic mean is often appropriate for the computation of loads of a contaminant, while a median or geometric mean is the preferred statistic for describing an 'average' concentration. For the arithmetic, geometric, and harmonic means the following is always true:

$$\text{harmonic mean} \leq \text{geometric mean} \leq \text{arithmetic mean}.$$

The three means are only equal if all the sample values are identical. Parkhurst (1998) discusses in more detail the relative merits of the geometric mean and arithmetic mean in the context of concentration data.

Variability is another extremely important characteristic of a distribution of results (see Table 6.3). The simplest measure of variability is the range — the difference between the largest score and the smallest score. The range is rarely used as a measure of variability because it is grossly affected by extremes in the data set (after all, it is defined as the difference between the two extremes). The most commonly used measure of variability is the variance or its (positive) square root, the standard deviation. The standard deviation is preferred because it has the same units of measurement as the original data. However, the variance is an important parameter in inferential statistics such as analysis of variance. One difficulty with the standard deviation is that it is not readily comparable among different populations or samples because it tends to be numerically higher as a result of an increasing mean. A dimensionless measure that overcomes this difficulty is the coefficient of variation (CV) which is defined as the ratio of the standard deviation and the mean.

Table 6.2. Measures of central tendency (adapted from Spiegel 1992)

Arithmetic mean	$\bar{X} = \sum_{i=1}^{n} X_i / n$ where X_i denotes the ith observation in a sample of n
α% trimmed mean	$\bar{X}_{T,\alpha}$ obtained by trimming (i.e. removing) α% off both ends of the ordered sample and computing $\bar{X}$ for the remainder. Used when outliers are present. α is typically 10% or 20%.
Mode	most frequently occurring value
Median	the middle value: half the values are numerically smaller and half are numerically larger
Geometric mean	the nth root of the product of n sample values (>0): GM = $(x_1 x_2 \ldots x_n)^{1/n}$. It is always less than the mean.
Harmonic mean	the reciprocal of the summation of n sample reciprocal values: HM = $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$

In addition to measures of location and dispersion (variability), another statistical quantity that has assumed greater significance and utility in water quality applications is the percentile of a distribution. For example, when examining physical–chemical stressors, the Water Quality Guidelines (ANZECC & ARMCANZ 2000) have adopted a triggering process based on a comparison of the 50th percentile at a test site with the 80th percentile at a reference site.

The pth percentile is the value that is greater than or equal to $p\%$ of all values in a distribution; e.g. 50% of all values in a distribution are numerically less than or equal to the 50th percentile (otherwise known as the median) while 80% of all values are numerically less than or equal to the 80th percentile. The 25th, 50th, and 75th percentiles are called the quartiles (denoted Q_1, Q_2 and Q_3) because they divide the distribution into four parts of equal probability.

Table 6.3. Common measures of variation (adapted from Spiegel 1992)

Range	(largest value) – (smallest value)
Interquartile range (IQR)	75th percentile – 25th percentile (percentiles and quartiles are defined in the text)
Sample variance (s^2)	$s^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$
Sample standard deviation (s)	The square root of the variance; it has the same units as the central tendency statistics
Windsorized standard deviation (s_T)	Used as a measure of spread of the trimmed mean $\bar{X}_T$ (Table 6.2). Obtained by replacing trimmed values, identified in computation of $\bar{X}_T$, with values that were next in line for trimming (one from each end), and computing the standard deviation s of this new sample; s_T is then obtained as $s_T = s\sqrt{(n-1)/(k-1)}$ where k is the size of the trimmed sample.
$\frac{\text{IQR}}{\text{median}}$	A dimensionless robust measure of spread
Coefficient of variation (CV)	The standard deviation divided by the mean; it is therefore dimensionless and can be used for comparisons among different samples

Table 6.4. Taxonomy of common types of graph and their applications

Graph type	EDA	Outlier detection	Distribution checking	Trend detection	Multivariable relationships	Model diagnostics	Process control	Cause and effect
Scatterplot (bi-plot)	✓	✓		✓	✓	✓		
Scatterplot matrix	✓	✓		✓	✓	✓		
Trellis graph	✓				✓			
Chart (line, bar, pie)	✓			✓				
Interval (error bars, confidence intervals)		✓				✓		
Histogram	✓	✓	✓			✓	✓	
Box-plot	✓	✓	✓			✓	✓	
Dot-plot	✓			✓				
Stem and leaf	✓	✓	✓			✓		
Time series	✓	✓		✓		✓	✓	
3D (contour, scatter, surface)	✓	✓		✓	✓			
Probability and Q-Q plots	✓	✓	✓					
Scatterplot smoothers (LOESS)	✓			✓				
Residual plot		✓	✓	✓		✓		
Run chart	✓			✓			✓	
Control chart (Xbar, CUSUM, EWMA, Range, MA)	✓						✓	
Pareto	✓						✓	
Fishbone	✓							✓

6.3.2. Data Visualisation

With recent advances in computer hardware and software, high quality and sophisticated graphics have become readily accessible. It is strongly recommended that relevant graphs of the data are drawn before any formal statistical treatment and analysis are done (Tufte 1983). Simple graphical devices such as histograms, box-plots, dot-plots, scatter plots, probability plots, etc., can assist statistical analyses and make it easier to interpret data with respect to:

- data anomalies and errors,
- outliers,
- properties of the distributions (location, dispersion, skewness),
- trends over time, space, attributes,
- relationships (existence of and type),
- checking assumptions appropriate to the distributions (e.g. normal probability plots),
- time series analysis,
- reducing the number of dimensions (visualising high dimensional data by projecting it into lower dimensions),
- operational performance (e.g. control charts).

A list of graph types and potential applications is given in Table 6.4. Most of these plots are available in the statistical software packages referred to in section 6.1.

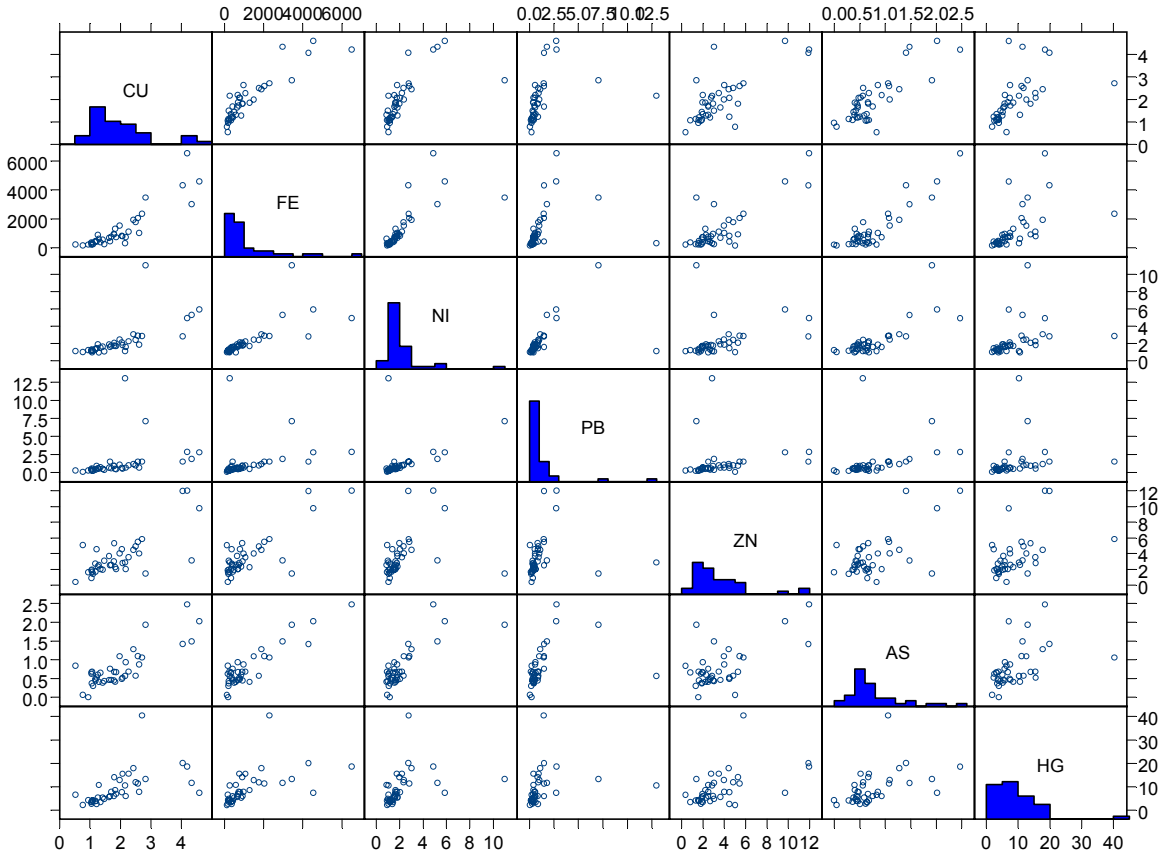

Figure 6.2. Scatterplot matrix for concentration of seven metals from a single water sample

For most water quality analyses, several different attributes or measures of water quality have to be assessed simultaneously. Important information can be overlooked if water quality variables are analysed one at a time. Relationships between individual variables can be detected relatively easily from graphical summaries such as the scatterplot matrix in Figure 6.2, which shows scatterplots and histograms for the concentrations of seven metals in water samples from a Victorian lake. A graphical display like this enables the analyst to assimilate the trends, relationships, and distributions of a number of variables that have been measured on the same water sample. Histograms for individual variables are shown on the diagonal while plots for pairs of variables are displayed on the off-diagonal (the labelling of axes is reversed to create two plots for each pair of variables on the off-diagonal).

6.3.3. Control Charting

Statistical Process Control (SPC) dates back to the 1930s and is deeply rooted in industrial applications where it is vitally important to control drift and variation in a process, to maintain production quality. Control charting techniques used for the last 70 years in industry have an important role to play in an environmental context. They are particularly relevant to water quality monitoring and assessment. Regulatory agencies are moving away from the 'command and control' mode of water quality monitoring, and recognising that, in monitoring, the data generated from environmental sampling are inherently 'noisy'. The data's occasional excursion beyond a notional guideline value may be a chance occurrence or may indicate a potential problem. This is precisely the situation that control charts target. They not only provide a visual display of an evolving process, but also offer 'early warning' of a shift in the process level (mean) or dispersion (variability). For further information, see Montgomery (1985) or Bennett and Franklin (1988 Chapter 10).

6.3.4. Data Coercion (Transformations)

Mathematical transformations of water quality data are usually undertaken with at least one of the following objectives in mind:

- to restore a greater degree of linearity between two or more variables,
- to stabilise the variance of some variable over time, space, or some other attribute,
- to restore a greater degree of normality in the distribution of some variable.

The identification of a suitable transformation (if it exists) is largely a trial and error process and will depend on the objectives of the exercise. Sometimes the theory or model that is being fitted to the data suggests the form of the mathematical transformation. For example, the analyst may suspect that the concentration of a nutrient is described by a power relationship of the form:

$$C = kQ^p \ ,$$

where C is concentration and Q is flow and k and p are unknown parameters. A simple logarithmic transformation yields a linear equation in $\log(C)$ and $\log(Q)$:

$$\log(C) = \log(k) + p\log(Q).$$

Thus, a log–log graph of the raw data will enable a quick assessment of the appropriateness of this model. Furthermore, estimates for $\log(k)$ and p are readily obtained as the intercept and slope respectively of the fitted regression line.

Another common goal when transforming data is to reduce the effect of a relationship between the mean and the variance. In a number of distributions (including the gamma and log-normal) the variance is related to the mean. Thus statistical tests designed to examine the equality of means will be affected by the non-constant variance. Some common transformations are provided in Table 6.5.

Table 6.5. Variance stabilising transformations (adapted from Ott 1984)

Mean (μ) ~ variance (σ^2) relationship	**Transformation**
$\sigma^2 = k\,\mu$ (Poisson data have $k = 1$)	$Y' = \sqrt{y}$ or $\sqrt{y + 0.375}$
$\sigma^2 = k\,\mu^2$	$Y' = \log(y)$ or $\log(y + 1)$
$\sigma^2 = k\,\pi(1 - \pi)$, $0 < \pi < 1$ (Binomial data have $k = 1/n$)	$Y' = \sin^{-1}(\sqrt{y})$

It is common practice to transform data when one or more assumptions of a proposed statistical test appear to have been violated. Many analysts transform data to try and restore some semblance of normality. In many instances this is unnecessary. A number of standard statistical procedures (such as ANOVA, *t*-tests, etc.) are relatively robust in the presence of slight to moderate departures from normality.

Rather than attempting to achieve normality, the analyst should ensure that the distribution of the data has a reasonable degree of symmetry. Significant distortions in the test results are only likely to occur in cases of high skewness and/or high kurtosis. It is far more important to check that data have homogeneous variances (i.e. the variances of the variables of interest are constant over different groups, times, or space) and are independent. Data that are either spatially or temporally correlated (or both) are not amenable to the statistical test procedures described in this document.

The identification of an appropriate transformation is often a matter of trial and error. However, within the class of power transformations, Box and Cox (1964, 1982) have developed a systematic procedure. The method seeks to identify a value of the transformation parameter, λ in the expression

$$Y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

such that values of the transformed data ($y^{(\lambda)}$) exhibit a greater degree of normality than the original set of y values. Note: this transformation is applicable to non-negative data only. The computations associated with this process are too complex to be described here, and since the method is computationally intensive, it is best handled by computer software such as MINITAB® or S-PLUS®. See further discussion in section A5.1.2.

6.3.5. Checking Distributional Assumptions

Many statistical methods of inference rely on the assumption that the sample data have been randomly selected from a larger population of values that is normally distributed. There are good reasons why the normal distribution enjoys such a prominent role in theoretical and practical statistics. First, many naturally occurring phenomena actually exhibit normal-shaped distributions. Second, the important Central Limit Theorem in statistics assures us that even when the distribution of individual observations is non-normal, aggregate quantities (such as the arithmetic mean) will tend to have a normal distribution. Another often-used, although less convincing argument for the use of normal-based inference is that the mathematics is 'tractable' and a large number of statistical procedures have been developed around the notion of random sampling from a normally-distributed population.

The properties of the normal distribution are well known and will not be repeated here. What is important is our ability to decide if a particular set of data can be assumed to have come from some population of values whose underlying distribution is normal (or some other specified form). Before the widespread availability of computers, plotting the data on probability paper provided a check of distributional assumptions. On probability paper the axis scales are specially constructed so that

when the distributional assumption is satisfied the resulting probability plot will be linear. This is akin to plotting data on log–log scales to check the assumption of a power relationship between two variables as was done in section 6.3.4. For more information on the use of probability plots see Ott (1984).

Many environmental variables are decidedly non-normal in their distributions. River flows may be normally distributed at a single locality for a short period of time, but on broader time scales they may decrease systematically with time in the absence of rainfall (time of sampling becomes a dominating influence); and, of course, a rainfall event will have a dramatic influence on flow.

Statistical software can do the computations associated with checking distributional assumptions. Most statistical software packages have the facility for testing the assumption of normality. Other packages, such as S-PLUS®, offer more flexibility in the types of distributions that can be examined. By way of examples, consider the distribution of arsenic concentrations obtained downstream from a gold mine (Figure 6.3), and Worked Example 1, page A5-28.

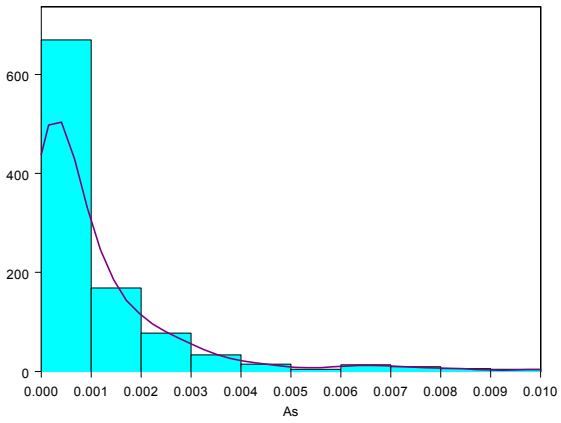

Figure 6.3. Downstream arsenic concentrations

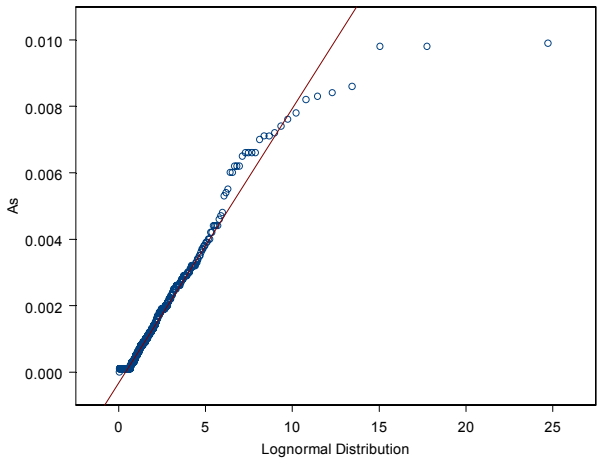

Figure 6.4. Log-normal probability plot for arsenic concentration data

Figure 6.3 reveals the non-normal shape of the data and it might be speculated that a *log-normal* distribution would be a more appropriate probability model. One way of checking the log-normal assumption would be to test the normality of the logarithms of the original arsenic data. A more direct approach is to inspect a log-normal probability plot (Figure 6.4). The tails of Figure 6.4 depart from linearity, suggesting that the data are not well described by a log-normal probability model and that a more suitable distribution may have to be sought.

6.3.6. Trend Detection

One of the principal objectives of water quality monitoring is to assess changes over time. In many instances this is driven by the need to compare water quality and a guideline value, although issues related to salinity, flow and climate variability are also important drivers.

A number of statistical methods are available to assist with trend analysis and these range from simple descriptive tools, such as time series plots, to more sophisticated modelling techniques that attempt to separate out a signal from 'noise'. These analyses have been greatly facilitated by the development of a variety of software tools. Figure 6.5a depicts arsenic time series data with the trend and 95%

prediction limits[2] overlaid. While the trend line shown is quite reasonable, care must be exercised when attempting to extrapolate beyond the range of the observed series. It is immediately apparent from Figure 6.5a that the arsenic concentrations increased quite dramatically in 1994. This could indicate altered mine practices or different ore-body characteristics and provides some explanation for the departure from linearity in the log-normal probability plot of Figure 6.4.

The trend exhibited in Figure 6.5a is revealing, but it is not the only feature of a time series with which we should be concerned. The variability of a process is equally as important as, and sometimes more important than real changes in level. Our definition of variance (Table 6.3) is based on departures from the mean (represented as the trend), so any assessment of variation should be performed on the de-trended data. The residuals (i.e. original observation minus trend) from this process can be plotted as a separate time series. This has been done for the arsenic data (Figure 6.5b).

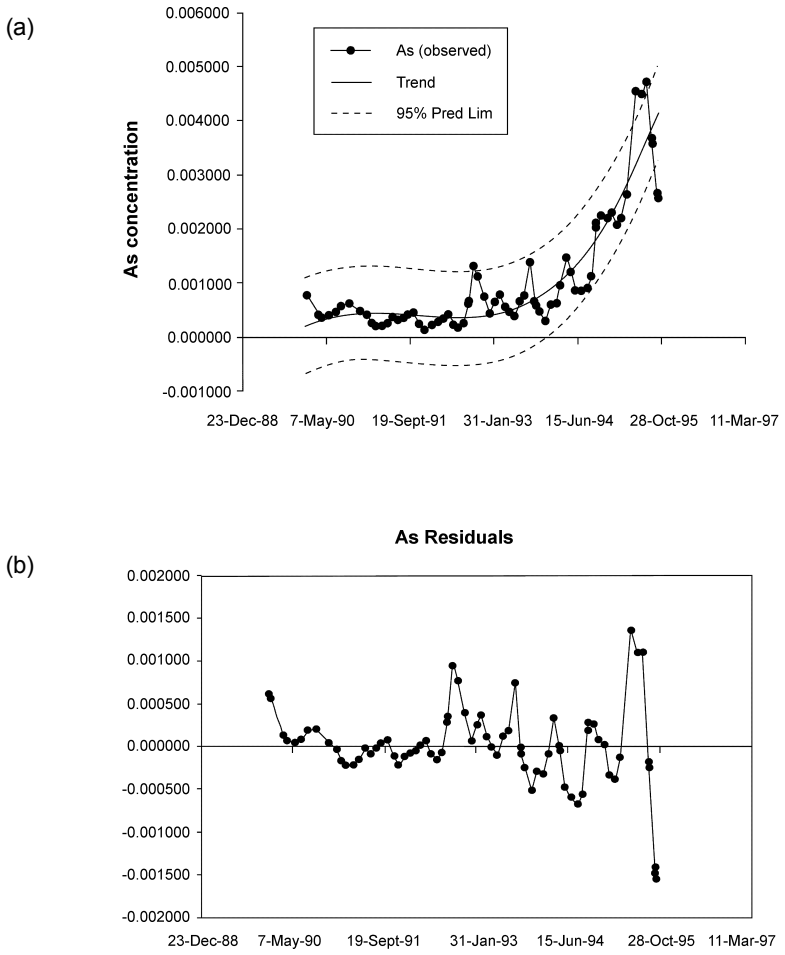

Figure 6.5. (a) Time series representation of the arsenic concentrations of Figure 6.3; (b) Time series plot of the arsenic residuals

[2] Prediction limits are used to establish bounds on an individual prediction whereas confidence limits place bounds on a mean. Since the variation associated with the prediction of a single value is greater than the variation associated with the estimation of a mean, the width of a prediction band will be greater than the width of a confidence band for some fixed certainty level.

Figure 6.5(b) shows an apparent change in the process variation beginning in 1993. Further inquiry might determine if this observation was the result of natural or human-induced effects.

6.3.7. Smoothing

Given the high variability of most natural ecosystem processes (or of indirect processes that influence them), it is not surprising that water quality data also exhibit a high degree of variation over both space and time. This high natural 'background' variation tends to mask trends in water quality parameters and reduces our ability to extract a signal from the 'noise'. Simple graphical tools such as scatterplots and time series plots can only provide a combined view of both the trend and the noise. However, so-called robust smoothers are techniques that 'let the data speak for themselves' and have been shown to be remarkably effective in teasing out a signal from very noisy data (Cleveland 1979).

Robust smoothers work by placing a 'window' over a portion of the data, computing some numerical quantity such as mean or median and then 'stepping' the window across the data and repeating the process. The collection of statistics obtained in this way can be plotted (Figure 6.6) at the mid-points of the intervals to obtain a smooth representation of the underlying process. Figure 6.6 reveals the emergence of a strong upward trend after 1994 on top of highly cyclical fluctuations. The amount of smoothing in these plots is controlled by the user who must specify a 'span' (i.e. the fraction of data to be captured in the moving window) and the number of passes over the data. Greater smoothing is achieved by increasing the span and/or the number of passes (Chambers and Hastie 1992).

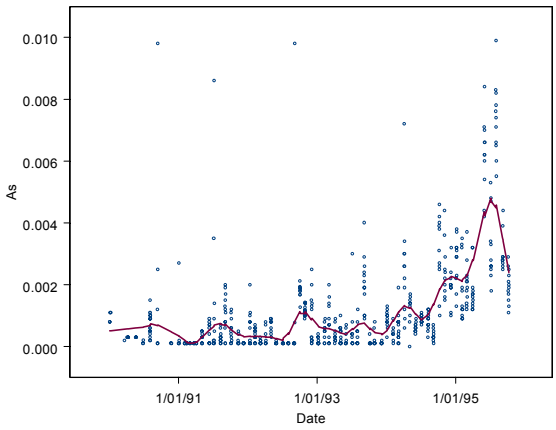

Figure 6.6. Smoothed arsenic data

6.4. Inference

The preceding sections of this chapter have focused on data analysis — that is, the process of summarising, presenting, and describing the information contained in sample information. While this is an important activity in its own right, most statistical analyses are concerned with inference — that is, methods for inferring something about some characteristic of a population of values, based on the limited information contained in a sample drawn from that population (see Table 6.9, page 6-26). In this context, statistical inference may be regarded as decision-making under uncertainty, and is thus imperfect.

6.4.1. Estimating an Unknown Water Quality Parameter

The true value of the concentration for phosphorus in a water storage is never known unless we drain the storage and measure all the phosphorus and the volume of water. Nevertheless, with an appropriate sampling regime, limits for the true value can be established. For example, the 95% confidence limits for a population mean give the range within which we can be 95% sure the true population value lies. To calculate a confidence interval for the true mean, the following formula can be used in conjunction with Table 6.6 (see also section A5.1.5.1),

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}},$$

where $\overline{X}$ is the sample mean, s is the sample standard deviation of n observations and $t_{\alpha/2}$ is a so-called 'critical' value from the t-distribution having degrees of freedom n–1.

Table 6.6. Critical *t*-values for selected confidence intervals and degrees of freedom (df)

df	90%	95%	99%	99.9%
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.599
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.768
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.690
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.659
30	1.697	2.042	2.750	3.646

For non-normal data, it is common to transform the data to yield approximate normality, then calculate confidence limits for the transformed data. It is possible to obtain approximate limits for the untransformed data by back-transformation.

Normality is not the only assumption made in using confidence limits. It is assumed that the measurements made on water samples are drawn at random from the water body in question, to protect against systematic bias. It is also assumed that the measurements are independent of each other. This might not be so if measurements are taken only from close to shore, as they will be more

similar to each other than would be measurements taken at random localities. Such measurements would be pseudoreplicates (Hurlbert 1984), and the confidence limits obtained would have an actual confidence level different to the assumed confidence level.

6.4.2. Testing Hypotheses

When a statistical hypothesis is being tested, the lack of complete information gives rise to Type I and Type II errors (Table 6.7); see section A5.1.5.2.

Statistical hypothesis tests are generally referred to as either parametric or nonparametric (i.e. 'distribution-free'). The distinction is that, in the former, the test procedure has been developed by assuming a parametric form for the underlying distribution of data values (e.g. normally distributed). The nonparametric tests relax these assumptions and are thus more robust. The price paid for this robustness is a reduction in statistical power when the assumptions of an equivalent parametric test would have been met.

It was remarked at the beginning of this chapter that the routine application of classical 'significance testing' procedures for water quality and ecological assessments is under scrutiny. Environmental researchers such as Underwood (1990, 1991, 1994), Fairweather (1991), and Green (1989, 1994) have helped raise the awareness of the need for proper statistical design and analysis in ecosystem assessment. Acronyms such as BACI, BACIP (BACI with temporal and paired spatial replication), and MBACI (BACI with full spatial and temporal replication) have helped increase this awareness. A cautionary note is nevertheless warranted. Much practical effort tends to be spent on activities that can distract the researcher from uncovering and modelling more important processes. It can be a difficult and sometimes futile exercise to attempt to identify an 'environmental control' in a landscape that has been disturbed and modified by human intervention. Similarly, the assumptions of ANOVA and related techniques have caused many analysts to think that that data must be coerced into normality at all costs.

Table 6.7. Types of error in statistical hypothesis testing

Decision	True state of nature	
	H_0 true	H_0 false
Accept H_0	✓	Type II error
Reject H_0	Type I error	✓

A balanced approach is required — one that acknowledges the rightful place of classical statistical inference yet encourages thinking and modelling outside the confines of simplistic, parametric analysis. Consistent with the recent trend away from strict 'compliance', data analysis should be focused more on discovering and understanding the spatial and temporal dynamics of an environmental impact, instead of on finding, for example, that the control and affected sites differ.

The Monitoring Guidelines cannot consider the entire range of statistical hypothesis tests in detail. Appendix 5 (e.g. sections A5.1.5.2, A5.1.6, etc.) explains the logic of significance tests and gives a more complete description of some commonly-used procedures. Two classes of statistical models that deserve greater attention in water quality studies are generalised linear models (McCullagh and Nelder 1983; Dobson 1990) and generalised additive models (Hastie and Tibshirani 1990).

Briefly, generalised linear models provide a far more flexible framework than 'conventional' methods (such as *t*-tests and ANOVA) for analysing data and making inferences. This is because of two main enhancements: (i) the error distribution is not confined to the normal probability model — i.e. log-normal, gamma, exponential, inverse Gaussian, binomial, Poisson and others are all permissible; and

(ii) generalised linear models accommodate non-linear relationships between the mean response and a set of predictor variables.

In a similar fashion, generalised additive models (or GAMs) have been devised to increase the flexibility of statistical modelling. Rather than imposing and estimating some predefined model, GAMs replace the usual linear function of an independent variable by an unspecified smooth function. In this sense, the model is nonparametric because a parametric form is not imposed on the functions — they are suggested by the data. For more discussion see section A5.1.12.

6.4.3. Comparison of Test Statistics and a Guideline or Trigger Value

This section discusses three techniques for comparing test statistics and water quality guidelines (see subsections 6.4.3.3–6.4.3.5 below). Water quality guidelines for fresh, marine, estuarine waters and groundwaters are provided in the revised Water Quality Guidelines document (ANZECC & ARMCANZ 2000), and that document also describes how the guidelines are derived and how they should be compared with test data. To give context to the discussion below, some of that background information is repeated here in sections 6.4.3.1 and 6.4.3.2.

For toxicants in water, test data are compared against default guideline values. The revised Water Quality Guidelines document (ANZECC & ARMCANZ 2000) recommends that the 95th percentile of concentration values at the test site should be less than the default guideline value for the toxicant (see Water Quality Guidelines section 7.4.4).

For biological measurement parameters and physical and chemical stressors, the revised Water Quality Guidelines advises the stakeholders for a waterbody to assess change in its quality (at its test site(s)) by comparing it with a relatively unimpacted and healthy reference waterbody or reference site(s). The reference waterbody and the test site(s) should be as similar as possible, with similar geology and climate (see Water Quality Guidelines section 3.1.4). Other matters to be considered when choosing a reference site are discussed in this Monitoring Guidelines document in section 3.2.1 and section 3.2.2, and in the Water Quality Guidelines section 3.1.4.

The manner in which test data for biological parameters are compared against reference data is discussed in the Water Quality Guidelines section 3.2.4. These issues are not discussed any further in this section.

The Water Quality Guidelines advocates that for physical and chemical (non-toxicant) parameters, the median quality values of fresh and marine waters should be lower than the 80th percentile of concentration values of a suitable reference site (above the 20th percentile for parameters such as dissolved oxygen where low values are the problem). Thus the 80th and 20th percentiles act as the trigger values (see the Water Quality Guidelines section 7.4.4).

6.4.3.1. Computation of Reference Percentiles and Their Use As Triggers

The notes in this subsection outline the derivation and use of percentiles of the reference site data for physical and chemical stressors, as described in the Water Quality Guidelines section 7.4.4. See that source for more detail.

The Water Quality Guidelines recommends that the computation of the 80th percentile at the reference site be based on the most recent 24 monthly observations there. The suggested procedure is as follows:

(i) arrange the 24 data values in ascending (i.e. lowest to highest) order; then

(ii) take the simple average (mean) of the 19th and 20th observations in this ordered set.

Each month, obtain a new reading at the reference (and test) sites. Append the reference site observation to the end of the original (i.e. unsorted) time sequence of reference site data, and then apply steps (i) and (ii) above to this new set of 24 data values. Note that even though only the most recent two years' data are used in the computations, no data are discarded.

A trigger for further investigation of the test waterbody will be deemed to have occurred when the *median* concentration of a particular measurement parameter in n independent samples taken at the test waterbody exceeds the 80th percentile (or is below the 20th percentile if 'less is worse') of the same measurement parameter at the reference site. A minimum of two years of consecutive monthly data at the reference site is required before a valid trigger value can be established based on that site's percentiles. If this requirement has not been satisfied, the median of the data values measured at the test site should be compared to the appropriate default guideline value identified in the Water Quality Guidelines.

The advantages of using a percentile of the reference distribution are:

- it avoids the need to specify an absolute quantity; and
- the trigger criterion is being constantly updated as the reference site is monitored, and therefore it reflects temporal trends and the effects of extraneous factors (e.g. climate variability, seasons).

Implementation of the trigger criterion is both flexible and adaptive. For example, the user can identify a level of routine sampling (through the specification of the sample size n) that provides an acceptable balance between cost of sampling and analysis and the risk of false triggering. The method also encourages the establishment and maintenance of long-term reference monitoring.

6.4.3.2. Number of Samples at the Test Site

The choice of number of samples (sometimes called 'sample size') to be taken at the test site is arbitrary, although there are implications for the rate of false triggering. For example, a minimum resource allocation would set $n = 1$ for the number of samples to be collected each month from the test site. If the distribution of reference site values is identical to the distribution of values measured at the test site, the chance of a single observation from the test site exceeding the 80th percentile (or less than the 20th percentile) of the reference distribution is precisely 20%. Thus the Type I error (the risk of triggering a false alarm) in this case is 20%. It can be reduced by increasing n. For example, when $n = 5$ the Type I error rate is approximately 0.05. The concomitant advantage of taking more samples is the reduction in Type II error (the risk of falsely claiming a degraded water body to be 'acceptable') (see the Water Quality Guidelines section 7.4.4).

The capability of a monitoring program that compares water quality in relation to a guideline value can be described in terms of the program's power curve (Ellis 1989; Ward *et al*. 1990; Donohue 2000 unpublished). In general, a power curve is used to establish the statistical performance of a test procedure. In the present context, values of the horizontal axis represent the degree to which the reference and test sites differ with respect to a measured parameter. The vertical scale gives the corresponding power — i.e. the probability that the test procedure will correctly identify a deterioration (at the test site) of a magnitude equal to the position on the horizontal scale. Section A5.1.10 in these Monitoring Guidelines discusses the power of a sampling scheme, and lists web sites that discuss or supply software with which power and the number of samples required for its achievement can be calculated.

6.4.3.3. Comparison between Test Data and Guideline Values By Control Charts

Control charts (see also section 6.3.3) can be usefully employed for comparing data values with a guideline or trigger value, and the Water Quality Guidelines recommends that they be used.

The advantages of control charts are that:

- minimal processing of data is required;
- they are graphical — trends, periodicities and other features are easily detected;
- they have early warning capability — the need for remedial action can be seen at an early stage.

When the monthly results for a test water body are graphed in a control chart they may look like Figure 6.7(a) (compared to a trigger obtained using the 80th percentile from reference site monitoring) or Figure 6.7(b) (compared to a single published guideline value). Both these diagrams appear in the Water Quality Guidelines section 7.4.4. Confidence limits can be added to each of the

plotted points when sample means are used, but this is not straightforward when dealing with percentiles (particularly a 'rolling' percentile at a reference site).

Confidence intervals themselves provide a good way of comparing observations of a water quality variable to a guideline value; see section A5.1.5.1

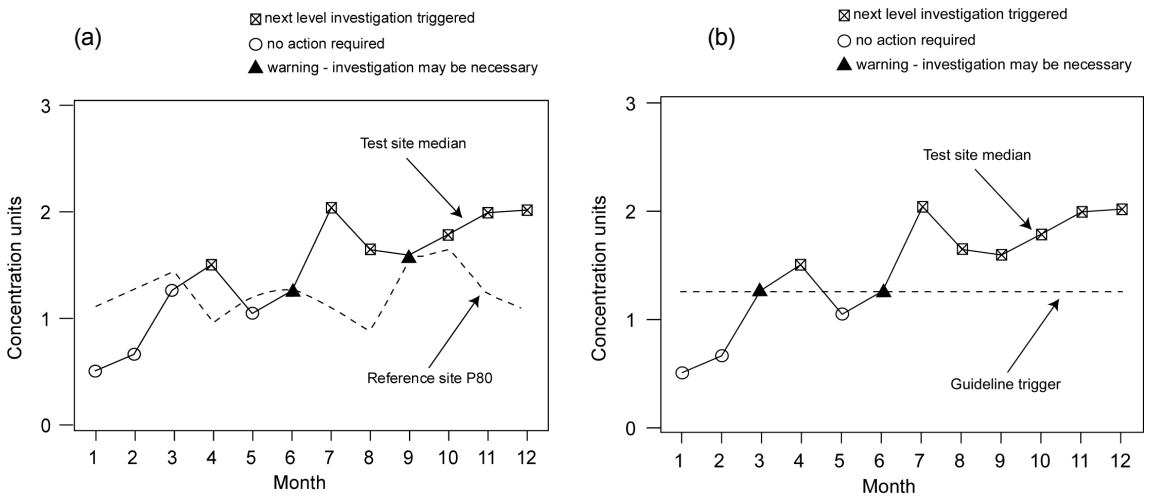

Figure 6.7. Control charts showing physical and chemical data (y-axis): (a) for test and reference sites plotted against time, and recommended actions; (b) for test site plotted against default trigger value and time, and recommended actions

6.4.3.4. Comparison between Test Data and Guideline Values by the Binomial Approach

The percentile-based method adopted by the Water Quality Guidelines for assessing water quality is a flexible nonparametric approach that is convenient and logical because inference can be made about extreme values (not just 'averages'). No assumptions have to be made about the distributional properties of the data obtained from either the test or reference sites.

A useful by-product of the percentile approach is that probabilistic assessments can be made about conformity to a guideline value without reference to the underlying statistical distribution of sample values. This requires an understanding of the binomial distribution at an introductory level.

When sampling is random and independent, variation in the number of sample excursions from a target follows the binomial distribution, regardless of the distribution of the raw data (Donohue 2000 unpublished). The binomial distribution describes the behaviour of the number of 'successes' any random process for which there are two mutually exclusive outcomes, such as heads/tails, higher than or less than. (See basic statistical texts, e.g. Walpole and Myers (1985), for details and formulae of the binomial distribution.)

Using the binomial distribution, the probability of obtaining *exactly* a certain number of samples with more (or less) than a certain level of a contaminant from a waterbody can be calculated. The binomial distribution is described completely by n, the number of random and independent samples collected, and by π, the assumed ecosystem rate of excursion. The probability of getting r samples that are worse than the trigger value ('sample excursions') from any set of n is (Ellis 1989):

$$\text{Prob.}(r) = \frac{n!}{(n-r)!r!}\pi^r(1-\pi)^{n-r} \quad (r = 0, 1, \ldots, n \;; \quad 0 < \pi < 1).$$

The binomial formula above gives probabilities for *individual* outcomes, that is probabilities associated with events of the kind 'exactly 3 out of 5 samples for which the guideline was exceeded'.

To assess the statistical significance associated with a particular outcome, a *cumulative* probability needs to be evaluated. The cumulative probability is defined as the probability of all events equal to *and more extreme than* the observed event. Thus, for example, if three out of five samples had a reading in excess of a guideline, then the appropriate cumulative probability is the probability of three or more samples (i.e. three, four or five) exceeding the guideline. This cumulative probability is known as the *p*-value in the language of hypothesis testing, and is conventionally compared to a nominal 0.05 level.

Example

Suppose the 90th percentile for cadmium in marine waters is not to exceed 5.5 µg/L. Five water samples at one location have the following Cd measurements (µg/L): 4.6, 4.4, 5.6, 4.7, 4.1.

Assuming that the 90th percentile for Cd in the marine waters is equal to 5.5 µg/L, then it can be seen that there is a 10% chance that a single reading will exceed this level. When n independent readings are taken, the probability of the number of samples r exceeding the guideline can be obtained from the binomial probability distribution. In this case (with $n = 5$ and $p = 0.1$), the expression is

$$\text{probability that } r \text{ out of 5 samples have Cd} > 5.5\ \mu\text{g/L} = \frac{5!}{r!(5-r)!}0.1^r 0.9^{5-r} \quad (r = 0,1,\ldots,5).$$

To compute a *p*-value for assessing the significance of our sample result, the probabilities of this event and those that are more extreme are computed and summed. In this case, the *p*-value is

$$\text{Prob.(1 exceedence)} + \text{Prob.(2 exceedences)} + \ldots + \text{Prob.(5 exceedences)}$$
$$= 1 - \text{Prob.(0 exceedences)} \quad = 1 - \frac{5!}{0!5!}0.1^0 0.9^{5-0} = 1 - 0.9^5 = 0.410.$$

Since this probability is much greater than the conventional 0.05 level of significance, there is insufficient evidence to assert that the guideline has been breached. In other words, it is quite probable that at least one out of five readings will exceed the 90th percentile. Note that this statement holds true irrespective of the actual numerical value of the 90th percentile or the distribution of readings. Most statistical software, including Microsoft EXCEL®, calculates binomial probabilities.

6.4.3.5. A Parametric Method for Comparing Test Data and Guideline Values

If a water quality parameter can be assumed to be normally distributed (perhaps after suitable transformation of the data), then the following procedure can be adopted (Fox 2000 unpublished).

Assume G, the guideline or trigger value, has been established so that in an 'undisturbed' system a proportion β of values will be less than G, with probability γ. Conformity to a guideline value will be demonstrated if the mean $\bar{X}$ of a sample of n readings satisfies

$$\bar{X} \le G - ks,$$

where s is the sample standard deviation and k is a factor depending on n. Values of k for selected n, β, and γ are provided by Fox (2000 unpublished) and reproduced in the table below.

Example

Referring back to the previous example, suppose it is known that the Cd levels in marine waters are normally distributed. The formula $\bar{X} \le G - ks$ uses information about the actual magnitude of the readings, not just their position relative to the 5.5 µg/L criterion. The assumption of normality is critical, as is the probability level, here called γ, because, as is shown below, conflicting results can be obtained.

For these data $\bar{X} = 4.68$ and $s = 0.563$. With $\beta = 0.90$ and $\gamma = 0.95$ and $n = 5$ we obtain $k = 3.407$ from the table below. The comparison with the guideline requires

$$\bar{X} \le 5.5 - (3.407)(0.563) = 3.582\ .$$

In this case, the sample mean of 4.68 exceeds 3.582 and, as in the example in 6.5.3.4, we are less than 95% confident that 90% of all Cd readings are below the required 5.5 µg/L.

n	$\gamma = 0.95$		$\gamma = 0.90$		$\gamma = 0.50$	
	$\beta = 0.95$	$\beta = 0.90$	$\beta = 0.95$	$\beta = 0.90$	$\beta = 0.95$	$\beta = 0.90$
2	22.261	20.583	13.090	10.253	2.339	1.784
3	7.656	6.156	5.312	4.258	1.939	1.498
4	5.144	4.162	3.957	3.188	1.830	1.419
5	4.203	3.407	3.400	2.742	1.779	1.382
6	3.708	3.006	3.092	2.494	1.751	1.361
7	3.400	2.756	2.894	2.333	1.732	1.347
8	3.188	2.582	2.754	2.219	1.719	1.337
9	3.032	2.454	2.650	2.133	1.709	1.330
10	2.911	2.355	2.568	2.066	1.702	1.324
15	2.566	2.068	2.329	1.867	1.681	1.309
20	2.396	1.926	2.208	1.765	1.671	1.301
30	2.220	1.777	2.080	1.672	1.662	1.295

However, with $\beta = 0.90$ and $\gamma = 0.50$ and n = 5 we obtain k = 1.382 from the table above. The comparison with the guideline requires:

$$\bar{X} \leq 5.5 - (1.382)(0.563) = 4.722 ,$$

which is satisfied by the observed sample mean of 4.68. This example highlights the importance of selecting the probability *a priori* — that is, *in advance of collecting the data*.

6.4.3.6. Further Discussion

For more discussion of the assessment of data against a quality threshold, see Miller and Ellis (1986), Gilbert (1987), Ellis (1989), Ward et al. (1990). Fox (2000 unpublished) and other papers are available on the CSIRO Environmental Projects Office web site, www.epo.csiro.au/library .

Donohue (2000 unpublished) is a paper that describes the binomial approach to assessing ecosystem excursions in relation to a trigger value, based on Ellis (1989). It includes a Western Australian case study in which the approach has been used. Copies can be obtained by contacting Rob Donohue, Environmental Officer, River and Estuary Investigation Section, Water and Rivers Commission, The Hyatt Centre, 3 Plain St, East Perth, WA 6004, phone (08) 9278 0586, fax (08) 9278 0532; robert.donohue@wrc.wa.gov.au.

6.5. Exploring Relationships

Relationships between pairs of water quality variables can be conveniently handled using the standard statistical tools of correlation and regression analyses.

6.5.1. Correlation Analysis

A useful adjunct to the scatterplot matrix presented in Figure 6.2 is a summary of the correlations between pairs of variables. The (Pearson) correlation coefficient is a numerical measure of the degree of linearity between two variables. Given two variables X and Y (where Y is notionally the dependent

variable and X the independent variable), the sample correlation coefficient (r) is computed using the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} .$$

Table 6.8. Broad interpretations of Pearson's correlation coefficient

Value of $\lvert r \rvert$	Interpretation
0.7 to 1.0	Strong linear association
0.5 to 0.7	Moderate linear association
0.3 to 0.5	Weak linear association
0 to 0.3	Little or no linear association

The correlation coefficient is constrained to lie in the interval $-1 \le r \le +1$. Broad interpretations of r are given in Table 6.8[3].

A correlation analysis is generally done before a more comprehensive regression analysis in which relationships between the variables are modelled and inferences are made about the likely values of true parameter values; see also section A5.1.11.

6.5.2. Regression Analysis

Usually, although not always, the objective of the regression analysis is to describe the relationship between a single dependent variable (Y) and a set of potential explanatory or independent variables (X_1, X_2, …, X_p). This is the multiple regression case. Simple linear regression refers to problems involving a single Y and a single X. Often the identification of independent and dependent variables is obvious. At other times there is no obvious labelling and an arbitrary choice can be made.

The term 'regression' covers a number of specific modelling approaches. These include:

- non-linear regression,
- multiple regression,
- stepwise regression,
- best subsets regression,
- robust regression.

An important assumption concerning the error terms of these regression models is that they are independent. Samples collected serially in time often display a degree of autocorrelation. For example, the concentration of phosphorus in storage at a particular time has a great bearing on the concentration an hour later, and probably a day later. If one measurement is well above the general trend, the other is likely to be also. Failure to ensure independence among measurements taken through time can have profound effects on the assumed Type I error rate, though the estimated parameters of the regression remain unbiased (Neter et al. 1996). One way to overcome temporal dependence is to select a sampling interval that is large enough to ensure no connection between consecutive measurements. Alternatively, various autoregressive models are available for analysing

[3] A related, and important statistic used to assess the adequacy of a fitted regression model is the coefficient of determination, R^2. In simple terms, R^2 is the proportion of total variation in some dependent variable that can be explained or accounted for by a regression model.

time series data, and the reader is referred to the text *Applied Linear Statistical Models* by Neter et al. (1996), and to Ott (1984) for an introduction.

6.5.3. Robust Regression

In terms of parameter estimation and identification of a good-fitting model**,** a significant drawback of the 'conventional' (ordinary least squares, OLS) method is its susceptibility to outliers in the data. The detection and handling of outliers was discussed in section 6.2.2 (see also section A5.1.3) and some broad recommendations were given. However, discard of observations is something that should not be undertaken lightly and needs to be fully justified. This poses a dilemma for subsequent statistical analysis given the potentially high leverage single aberrant observations may exert on the regression results. It is with these considerations in mind that alternative regression techniques have been devised which have been shown to be particularly resilient to data abnormalities. These techniques will give results which are in close agreement with classical (OLS) methods when the usual assumptions are satisfied, but differ significantly from the least-squares fit when the errors do not satisfy the normality conditions or when the data contain outliers. The statistical software package S-PLUS® has a rich suite of robust regression tools available. It includes the new Robust MM Regression, based on Rousseeuw and Yohai (1984), Yohai et al. (1991), and Yohai and Zamar (1998) as well as Least Trimmed Squares (LTS), Least Median Squares (LMS), and least absolute deviations (L1). See section A5.1.11 for more discussion.

6.5.4. High-Dimensional Data

Water quality monitoring programs often generate high-dimensional data that pose considerable challenges for the data analyst. Statistical analogues of many of the univariate techniques identified in Table 6.9(a) are available and these come under the statistical umbrella of multivariate analysis. These techniques rely heavily on a good understanding of more advanced statistical concepts and linear algebra (vectors and matrices). All the statistical software packages identified earlier in this chapter have multivariate statistical analysis capabilities. Before applying these more advanced statistical tools, the analyst should explore the data at hand and attempt to identify relationships between pairs and groups of parameters. Visualisation of data is necessarily limited to three dimensions, albeit as a two-dimensional projection. To overcome this 'curse of dimensionality', a number of statistical procedures have been devised in an attempt to reduce the number of dimensions with minimal loss of information contained in the original set of variables. One such technique that is potentially useful in water quality studies is Principal Component Analysis. The interested reader shuold consult any of the texts available on multivariate statistics.

Principal Component Analysis (PCA) constructs linear combinations of the original variables such that the resulting combinations account for the maximal amount of variation in the original data using considerably fewer constructed variables. The drawback is that the constructed variables (the 'components') generally have no real or physical meaning and have no recognisable units of measurement. However, as an exploratory analysis technique and dimension–reduction device, PCA has a potentially valuable role to play.

6.6. Changes in Space and Time

6.6.1. Time Series Analysis

Formal statistical analysis of time series data can be rather complex and requires a good degree of skill in identifying suitable models. Numerous texts have been written on time series analysis, although Cryer (1986) provides a good overview at an introductory level.

It has been previously remarked in this section that violation of the independence assumption can cause considerable distortion of the results of some standard methods of inference such as *t*-tests and ANOVA. Time series data tend to exhibit varying degrees of serial dependence. A particularly useful tool in identifying and characterising this serial dependence is the autocorrelation function or ACF. The ACF is the correlation between pairs of observations separated by a constant lag or time-step.

Further discussion of aspects of time series analysis and illustrative examples are in section A5.1.4.

6.6.2. Testing for Trend

A nonparametric test of trend that is often used in water quality studies is the seasonal Kendall test (Gilbert 1987; Helsel and Hirsch 1992).

An assumption of the trend test is that the trends are monotonic; that is, they consistently increase or decrease (Helsel and Hirsch 1992). If concentrations vary non-monotonically over the period being analysed, the results of linear tests for trend may be misleading (Robson and Neal 1996). Furthermore, it is assumed that observations are independent. When applied to data series that are not independent (that is, exhibit autocorrelation) the risk of falsely detecting a trend is increased (Ward et al. 1990; Esterby 1996). As a general rule, the level of serial correlation in a data series increases as the frequency of sampling increases. The maximum sampling frequency possible without encountering serial correlation can be thought of as the point of information saturation (Ward et al. 1990).

The seasonal Kendall test is based on the following statistic:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \text{sign}(x_j - x_k) \qquad \text{where} \quad \text{sign}(x_j - x_k) = \begin{cases} 1 & \text{if } x_j - x_k > 0 \\ 0 & \text{if } x_j - x_k = 0 \\ -1 & \text{if } x_j - x_k < 0 \end{cases}$$

which is then compared with a tabulated critical value to assess its significance.

While the seasonal Kendall test is robust under a variety of conditions, it is nevertheless important to examine the data for outliers (see section 6.2.2) and to adjust the data for the effect of nuisance variables. For example, many water quality indicators are correlated with flow and thus the seasonal Kendall test should be applied to the residuals (i.e. flow-corrected observations). These residuals can be obtained either as data predicted, where predicted values are obtained by explicitly modelling the relationship between flow and the variable of interest (e.g. using a regression model), or by using a nonparametric smoother.

6.6.3. Multidimensional Scaling

Multidimensional scaling (MDS) is a statistical technique that attempts to reveal relationships or patterns among a large number of variables by reconstructing 'similarities' between pairs of these variables in a 'reduced' space (i.e. one of fewer dimensions); see also section A5.1.14. Biologists and ecologists have found MDS analysis particularly useful in teasing out complex space–time interactions among biological variables. Like PCA, MDS also has some limitations, including the presentation of results in an abstract space, uncertainty surrounding the appropriate dimension of the reduced space, and a multitude of similarity measures on which to base the MDS. The calculations underpinning MDS are highly complex and are iterative — that is, a terminal solution is generally found only after a number of alternatives have been explored. The lack of agreement between the similarities or distances in the final MDS representation and the original input data is measured by the so-called 'stress statistic'. The derivation of fitted distances depends also on the type of MDS that is performed. There are numerous MDS procedures, although the main dichotomy differentiates between metric MDS and non-metric MDS depending on the measurement level of the data. Metric MDS assumes the input data are quantitative (i.e. measured on an interval or ratio scale) while non-

metric MDS assumes the data are qualitative (i.e. measured on a nominal or ordinal scale)[4]. There are differing views about the extent to which MDS can be used as an inferential tool. Computationally intensive methods are available that enable the researcher to conduct formal tests of significance, although the strength of MDS for water quality studies is more likely to reside in its ability to let the analyst discern patterns and trends visually.

6.7. Interpretation

After the data analysis, the monitoring team collates the results into a concise statistical summary, and assesses these results by use of residual diagnostics (see section A5.2 Worked Examples). This is the stage at which the team interprets the information the results provide, in the context of the objectives or questions the program was originally set up to answer. Interpretations might be, for example, that the values of a contaminant exceed the guidelines for ambient water quality because of the release of effluent by a sewerage system; or that the values of important measurement parameters before and after the building of a coastal recreation development differ significantly; or that two tested factors are not significantly reducing groundwater quality.

Once the monitoring team has expressed the interpretation concisely in non-technical terms, it can decide whether or not the program objectives have been satisfied, and whether it is appropriate to report to stakeholders.

If the interpretation does not support the conceptual model or the objectives have not been met, the model needs to be refined, the study needs to be redesigned, and new or additional data need to be collected and the analysis restarted.

To assist in the evaluation of detailed monitoring programs, the study team may consider seeking an independent peer review that can assess the program design and outcomes against the monitoring program objectives.

[4] See Appendix 5, section A5.1.1, for a discussion of measurement levels.

Table 6.9(a). Summary of common statistical procedures and applications for one-variable studies

SINGLE PARAMETER INFERENCE (adapted from Kitchens 1987)				
Parameter	**Application**	**Confidence interval**	**Hypothesis**	**Test statistic**
μ (population mean)	Inference about the 'average' value of a single water quality variable. Assumes a symmetrical distribution whose tails are not excessively long	Large sample $\bar{x} \pm z_{\alpha/2} s/\sqrt{n}$	H_0: $\mu = \mu_0$	$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
	Inference about the 'average' value of a single water quality variable. Assumes normally distributed data.	Small sample $\bar{x} \pm t_{\alpha/2,n-1} s/\sqrt{n}$ degrees of freedom = $n-1$		$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
	Inference about the 'average' value of a single water quality variable. Suitable for use with symmetrical distributions with long tails.	$\bar{x}_T \pm z_{\alpha/2} s_T/\sqrt{k}$ $\bar{x}_T$ is the trimmed mean (see Table 6.2); s_T is the standard deviation of the Windsorized sample (see Table 6.3) and k is the size of the trimmed sample.		$Z = \frac{\bar{x}_T - \mu_0}{s_T/\sqrt{k}}$
$\tilde{\mu}$ (population median)	Inference about the median value of a single water quality variable. For use with skewed distributions.	Large sample ($n > 20$): 1. Arrange data in ascending order. 2. Compute $C = (n - z_{\alpha/2}\sqrt{n})/2$ and reduce to next whole number. 3. The two limits for the confidence interval are identified as the C th observations from the low end and high end of the ordered data.	H_0: $\tilde{\mu} = \tilde{\mu}_0$	$Z = \frac{2T - n}{\sqrt{n}}$ where T = no. of observations > $\tilde{\mu}_0$
		Small sample ($n < 20$): Procedure as above for large samples, except C determined from Table A5.4.		Compute p-value directly using: $p = 1 - 0.5^n \sum_{t=0}^{T-1} \binom{n}{t}$ where T = no. observations > $\tilde{\mu}_0$
π (true population proportion)	Inference about a proportion (e.g. % conformity with a target, % 'defects' in a sample). Note: p in the formulae should be a fraction between 0 and 1.	$p \pm z_{\alpha/2}\sqrt{p(1-p)/n}$	H_0: $\pi = \pi_0$	$Z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$
σ^2 (population variance)	Inference about the variability of a single water quality variable. Assumes normally distributed data.	$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$ degrees of freedom = $n-1$	H_0: $\sigma^2 = \sigma_0^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
ρ_0 (true correlation coefficient)	Inference about the correlation between two water quality variables. Assumes normally distributed data.		H_0: $\rho = \rho_0$	$Z = \frac{\sqrt{n-3}}{2} \ln\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right]$

Table 6.9(b). Summary of common statistical procedures for two-variable studies

TWO PARAMETER INFERENCE (adapted from Kitchens 1987)				
Parameter	**Application**	**Confidence interval**	**Hypothesis**	**Test statistic**
μ_d (population mean difference)	Inference about the 'average' difference between pairs of a water quality variables. Assumes samples are related (e.g. 'before' and 'after') and differences are normally distributed.	$\bar{d} \pm t_{\alpha/2,n-1}\, s_d / \sqrt{n}$	$H_0: \mu_d = 0$	$t = \dfrac{\bar{d}}{s_d / \sqrt{n}}$
	Inference about the 'average' difference between pairs of water quality variables. No distributional assumptions required.		Wilcoxon signed rank test	$t = \dfrac{\bar{r}}{s_r / \sqrt{n}}$
$\mu_1 - \mu_2$	Inference concerning the difference between the means of two independent populations. Symmetrical distributions whose tails are not excessively long.	Large sample confidence interval. $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	$H_0: \mu_1 = \mu_2$	$z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
	Inference concerning the difference between the means of two independent populations. Normal distributions with equal variances.	Small sample confidence interval. $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\ s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$ where $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$ and degrees of freedom $= n_1 + n_2 - 2$		$t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$
	Inference concerning the difference between the means of two independent populations. Symmetrical distributions with long tails.	Large sample confidence interval. $(\bar{x}_{T_1} - \bar{x}_{T_2}) \pm z_{\alpha/2} \sqrt{\dfrac{s_{T_1}^2}{k_1} + \dfrac{s_{T_2}^2}{k_2}}$ $\bar{x}_T$ is the trimmed mean (Table 6.2); s_T is the standard deviation of the Windsorized sample (Table 6.3) and k is the size of the trimmed sample.		$z = \dfrac{\bar{x}_{T_1} - \bar{x}_{T_2}}{\sqrt{\dfrac{s_{T_1}^2}{k_1} + \dfrac{s_{T_2}^2}{k_2}}}$
$\theta_1 - \theta_2$	Inference about difference between two medians. No assumptions made about shape of distributions other than they are reasonably similar.	Wilcoxon Rank Sum test (this is an ordinary *t*-test applied to the rank-transformed data, i.e. individual data values replaced by their rank, r in the respective samples). Samples of at least 10 from both populations should be available for analysis.	$H_0: \theta_1 = \theta_2$	$t = \dfrac{\bar{r}_1 - \bar{r}_2}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ degrees of freedom $= n_1 + n_2 - 2$

Table 6.9b continued

$\pi_1 - \pi_2$	Inference about difference between two proportions (e.g. % conformity at two sites or times). Note: *p* in the formulae should be a fraction between 0 and 1.	$(p_1 - p_2) \pm z_{a/2}\ \mathrm{SE}(p_1 - p_2)$ where $\mathrm{SE}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	H_0: $p_1 = p_2$	$z = \frac{p_1 - p_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ and $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$
$\frac{\sigma_1^2}{\sigma_2^2}$	Inference about the equality of two variances. Assumes normally distributed data. *NB: The quotient* σ_1^2 / σ_2^2 *should be expressed such that the* ***larger*** *of the two (sample) variances is the numerator.*	$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2,\nu_1,\nu_2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2,\nu_2,\nu_1}$ **NB:** note the reversal of the degrees of freedom for the critical *f* value in the left and right sides of this expression; degrees of freedom ν_1 and ν_2.	H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$	$f = \frac{s_1^2}{s_2^2}$ critical F values are $F_{1-\alpha,\nu_1,\nu_2}$ for a 'less than' alternative hypothesis; F_{α,ν_1,ν_2} for a 'greater than' alternative hypothesis; and $F_{1-\alpha/2,\nu_1,\nu_2}$ and $F_{\alpha/2,\nu_1,\nu_2}$ for a 'not equal to' alternative hypothesis.