Appendix 5

Statistical Methods for Water Quality Monitoring Programs

A5.1. Application of Statistical Procedures in Chapter 6

The purpose of this appendix is to provide additional material and worked examples to clarify the application of the statistical procedures discussed in Chapter 6, and to assist with interpretation of statistical output in the context of water quality sampling, monitoring and assessment. The reader seeking more detail and discussion should consult appropriate references cited here and in Chapter 6.

A5.1.1. Summarising Data

Tables 6.2 and 6.3 in Chapter 6 identify a number of common statistical measures for describing the 'average' and 'spread' respectively of a distribution. However, not all summary statistics are appropriate to all types of measurement. The arithmetic mean is most appropriate for data measured on the interval and ratio scales. The median can also be calculated at the ordinal scale, while the mode can be determined at any measurement scale. Table A5.1 lists the most common summary statistics and the measurement levels required for their determination.

Table A5.1. The applicability of various summary techniques to data measured at each of four levels of measurement

Data reduction method	**Data type**			
	Nominal	Ordinal	Interval	Ratio
Frequency tabulations	✓	✓	✓	✓
Bargraphs	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Evenness Index	✓	✓	✓	✓
Median	✗	✓	✓	✓
Quartiles and percentiles	✗	✓	✓	✓
Interquartile range	✗	✗	✓	✓
Histograms	✗	✗	✓	✓
Frequency polygons	✗	✗	✓	✓
Arithmetic mean	✗	✗	✓	✓
Standard deviation	✗	✗	✓	✓
Variance	✗	✗	✓	✓
Coefficient of variation	✗	✗	✗	✓

- Nominal data are measurements assigned to one of several classes. These measurements are simply counts. The order in which the classes are presented is quite arbitrary, e.g. sex (male or female), maturity status (adult or juvenile), fish species assemblage (Carp, Bass, Perch, Murray Cod, or Redfin).

- Ordinal data are measurements with all the properties of nominal data, but these classes can be ranked in some order, e.g. algae at each site can be absent, sparse, common, abundant, or very abundant. We can say one measurement is larger than another, but we cannot say by how much.
- Interval data are measurements with all the properties of ordinal data; they can also be used to determine by how much one measurement differs from another. Values are measured with respect to an arbitrary zero. Temperature measured in degrees Celsius is an example where comparisons relate to an arbitrary zero, the freezing point of water.
- Ratio data are measurements with all the properties of interval data, but the values are measured with respect to a true zero. An example is phosphorus concentration; a zero means there are no phosphorus molecules present.

This classification is due to Stevens (1946).

A number of additional statistical measures are available to describe features other than location and dispersion. The *skewness* coefficient is a measure of a distribution's asymmetry while kurtosis is a measure of 'peakedness' — usually relative to the normal distribution. Symmetrical distributions such as the uniform and normal have zero skewness coefficients. Most statistical software packages have the facility for computing these quantities and formulae are available from statistical texts. Some important theoretical probability models encountered in water quality monitoring are the normal, gamma, and log-normal distributions. The gamma and log-normal distributions are always positively skewed (i.e. their long tail is to the right) — a feature commonly exhibited by water quality data. Examples of skewness and its impact on common measures of location are shown in Figure A5.1(a–c).

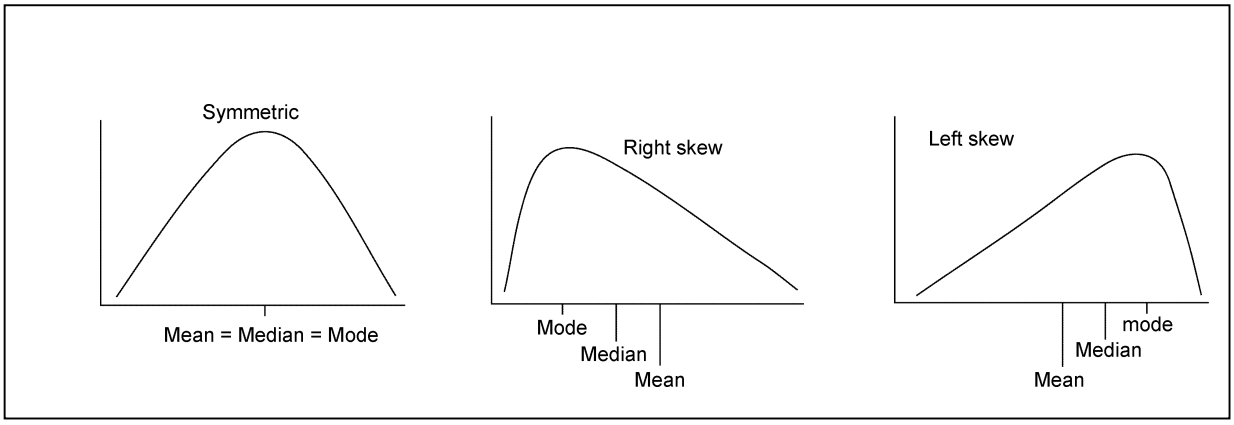


Figure A5.1. Comparison of three distributions and their influence on the measures of central tendency

A histogram of nitrate levels in a water body is shown in Figure A5.2 together with a smoothed estimate of the underlying distribution. The similarity to the gamma distribution (Figure A5.1(b)) and the log-normal (Figure A5.1(c)) distribution is apparent. Histograms are widely used to examine the basic shape and features of a set of data. As Figure A5.2 illustrates, many software packages provide the added facility of using robust smoothing techniques to overlay a 'density estimate' — essentially an estimate of the true or underlying distribution for the entire population of values. Another useful graphical device that permits the extraction of somewhat more quantitative information is the box-plot. The box-plot for the nitrate data is shown in Figure A5.3.

There are a number of ways of constructing and depicting the box-plot, although the essential features are similar. The width of the box typically represents the interquartile range (IQR); the 'whiskers' (lines extending either side of the box) span the range of the data; the horizontal line within the box

depicts the position of the median while the horizontal lines beyond the whiskers denote 'outliers'. Another useful facility offered by most software tools is the ability to generate separate box-plots based on the values of some other factor. For example, the nitrate data depicted in Figure A5.2 are, in reality, data from nine individual sites. Figure A5.4 shows the individual site box-plots. Sites A and B were from one location, sites C, D and H from a second location, and sites E, F, G and I from a third location. These groupings are revealed in Figure A5.4.

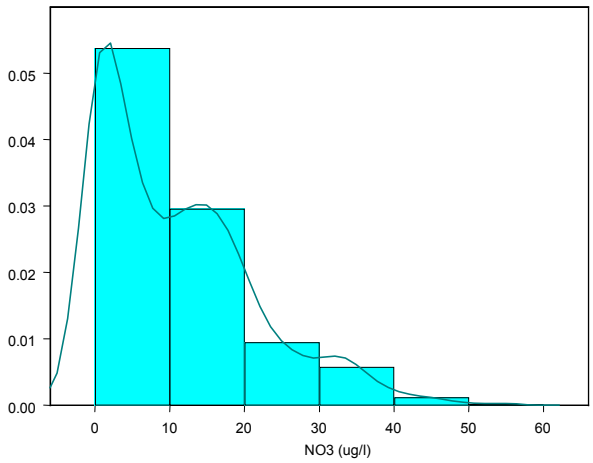


Figure A5.2. Sample histogram of probability densities (y-axis) for nitrate concentrations (µg/L), with smoothed distribution overlaid

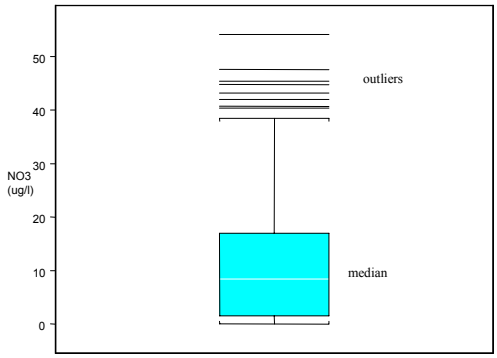


Figure A5.3. Box-plot for nitrate data

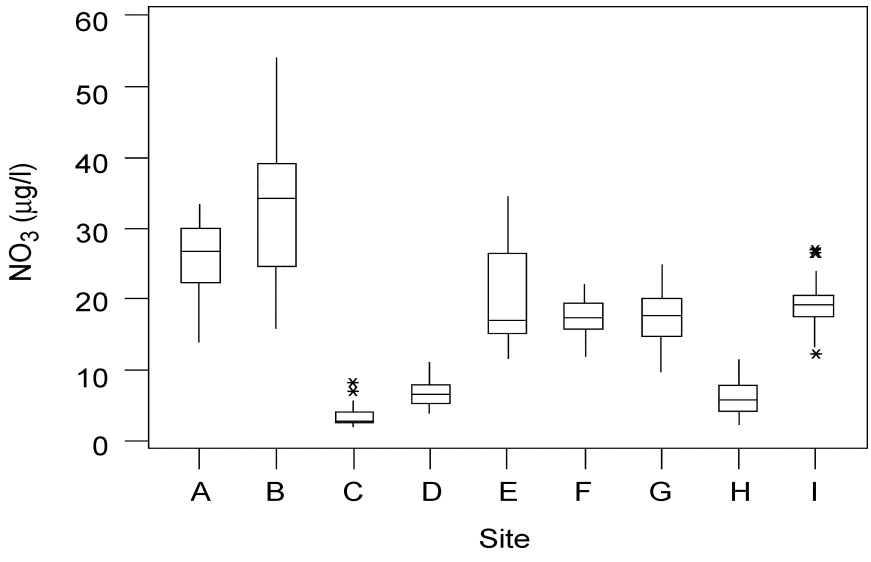


Figure A5.4. Site-specific box-plots for nitrate data (asterisks denote outliers)

Box A5.1. Illustrative panel about summarising data

The nitrate data displayed in Figure A5.2 are investigated in more detail using some standard statistical tools available in most statistical software packages.

Descriptive Statistics

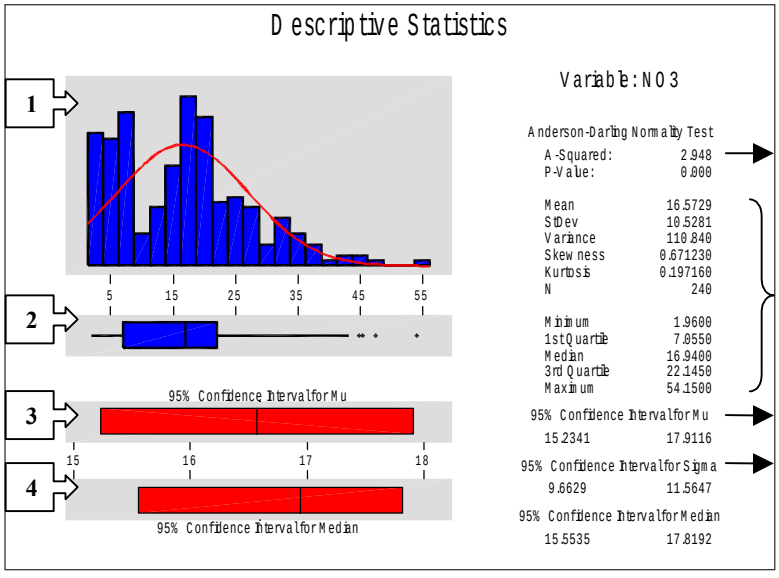
Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
NO3	240	16.573	16.940	15.955	10.528	0.680

Variable	Min	Max	Q1	Q3
NO3	1.960	54.150	7.055	22.145

Explanation:

N — sample size (240 in this case)
Mean — the *arithmetic* mean
Median — 50% of the data have values greater (less) than this value
Tr Mean — a *trimmed* mean obtained by averaging the middle 90% of data (this measure is less susceptible to the influences of extremes and/or outliers in the sample).
StDev —the sample standard deviation
SE Mean — the standard error of the mean. This statistic is a measure of the precision of the arithmetic mean ($SE = \frac{\sigma}{\sqrt{n}}$, where σ is the population StDev).
Min — the minimum value in the data set.
Max — the maximum value in the data set.
Q1 — the first quartile or equivalently, the 25th percentile.
Q3 — the third quartile or equivalently, the 75th percentile.

NB: The inter-quartile range (not shown) is equal to **Q3 – Q1**.



Results of a test of the data's normality. The low *p*-value here suggests the assumption of an underlying normal distribution is untenable.

Descriptive statistics for the data — essentially the same information as presented above.

Numerical information to be read in conjunction with the plot to the left. 95% confidence interval for the *true* standard deviation (see section 6.4.1 and section A5.1.5.1 for discussion of confidence intervals).

1. Histogram of data with smooth estimate of distribution overlaid
2. Box-plot of data
3. 95% confidence interval for the *true* (population) mean
4. 95% confidence interval for the *true* (population) median

A5.1.2. Transformations to Normality

The issue of transforming data prior to statistical analysis is covered in section 6.3.4 in Chapter 6. As noted there, the 'correct' application of many statistical procedures relies on the assumption that the data have been sampled from a parent population of values that are normally distributed. In many instances in water quality sampling, this assumption cannot be made. The Box–Cox transformation is often used in an attempt to restore some semblance of normality to non-normal data. The mathematical form of this transformation is given in section 6.3.4. The identification of the transformation parameter λ is greatly facilitated by a diagnostic plot similar to that shown in Figure A5.5.

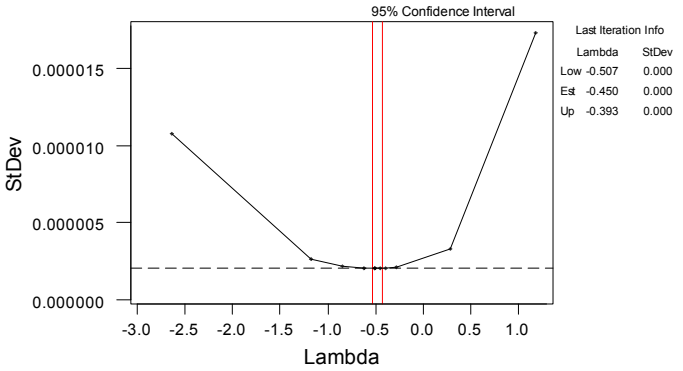


Figure A5.5. Diagnostic plot for the Box–Cox transformation

The optimal value for λ corresponds to the minimum ordinate value in the diagnostic plot of Figure A5.5. The vertical lines define a confidence interval for the true (but unknown) transformation parameter. From the printout on the right-hand side of the plot, we see that the best value for λ is –0.450. Thus, the best power transformation of this data (to improve normality) is

$$Y' = \frac{y^{-0.45} - 1}{-0.45}.$$

Figure A5.6 shows the histogram of the transformed data[1]; the *bimodality* of the transformed data is evident. This is suggestive of two independent processes operating, or else a breakpoint in the data.

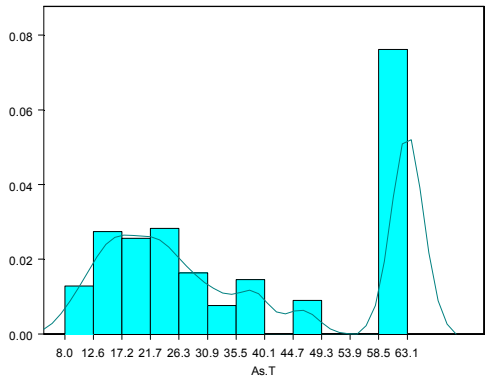


Figure A5.6. Histogram of transformed arsenic data

[1] Note. The software package used in this example has transformed the data using the relationship $Y' = y^{\lambda}$.

A5.1.3. Outlier Detection

The issue of data integrity is covered in section 6.2.2 in Chapter 6. Identification of 'unusual' observations or outliers is reasonably straightforward for single variables. However, as pointed out in section 6.2.2, the task in a multivariate context can be considerably more complex. An example of the joint relationship between two variables is shown in Figure A5.7. The univariate concept of a confidence interval is readily extended to a confidence ellipse in two or more dimensions. The projection of this ellipse onto the axes defines the confidence interval for the respective variable or parameter. The confidence ellipse allows us to make simultaneous inference about all pairs of values (Y_1, Y_2) jointly. For example, we can assert that we are 95% confident that all pairs (Y_1, Y_2) lie within the boundary of the joint confidence ellipse. Thus point P_4 in Figure A5.7 is an 'outlier' in the joint sense, although the one-at-a-time analysis would suggest that this point is not unusual with respect to either Y_1 or Y_2. An analysis of the points in Figure A5.7 is given in Table A5.2.

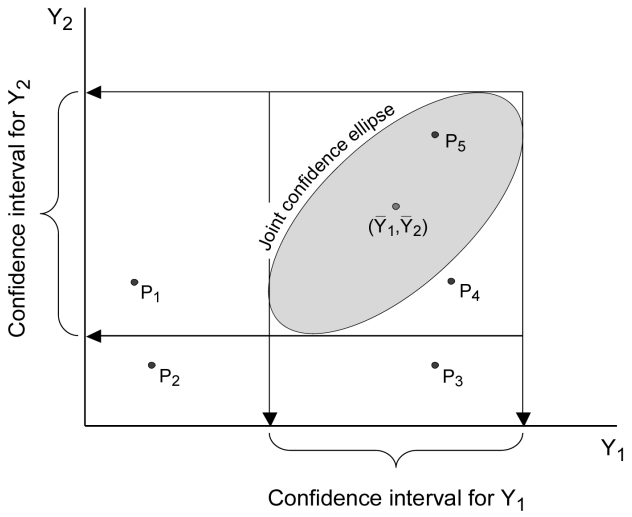


Figure A5.7. Relationship between univariate confidence intervals and the joint confidence ellipse for two variables, Y_1 and Y_2.

The computation of joint confidence ellipses is beyond the scope of this document. The reader interested in learning more about this topic is advised to consult a standard reference on multivariate statistics (e.g. Chatfield 1980).

Table A5.2. Outlier analysis of points in Figure A5.7

Observation	Y_1 outlier?	Y_2 outlier?	Joint {Y_1, Y_1} outlier?
P_1	✓	✗	✓
P_2	✓	✓	✓
P_3	✗	✓	✓
P_4	✗	✗	✓
P_5	✗	✗	✗

A5.1.4. Time Series Analysis

This section on time series analysis complements section 6.3.6 and section 6.6.1 in Chapter 6.

Formal statistical analysis of time series data can be rather complex and requires a good degree of skill in identifying suitable models. Some basic strategies are illustrated with application to the analysis of arsenic in section 6.3.6. Numerous texts have been written on time series analysis, although Cryer (1986) provides a good overview at an intermediate level.

In addition to the assumption of normality, many statistical tests further assume that the data are statistically independent. Violation of this assumption is potentially more serious than violation of the normality assumption and can cause considerable distortion of test results. By their very nature, time series data tend to exhibit varying degrees of serial dependence. A particularly useful tool in identifying and characterising this serial dependency is the autocorrelation function (ACF). The ACF is the correlation between pairs of observations separated by a constant lag or time-step. A plot of the ACF for the arsenic residuals of Figure 6.5b is shown in Figure A5.8.

There are a number of elements to this plot. The vertical lines represent the correlation at successive lags. The envelope represented by the dashed lines is a 95% confidence interval for the true autocorrelation. The simple interpretation is that vertical bars that extend beyond the envelope can be considered to be non-zero, while correlations that are contained within the envelope are indicative of non-significant (i.e. zero) correlations. From Figure A5.8 significant correlations are evident, out to about 15 time lags. The other interesting feature exhibited by Figure A5.8 is the slow linear decay of the ACF. This is suggestive of some structure in the residuals that has not been accounted for by the removal of trend alone. Further insights may be gained from an inspection of the partial ACF or PACF. This is similar to the ACF except that the correlation at lag k, say, is computed in such a way that it is independent of all intervening correlations (i.e. correlations from lags 1 through to k–1). The PACF for the arsenic residuals is shown in Figure A5.9.

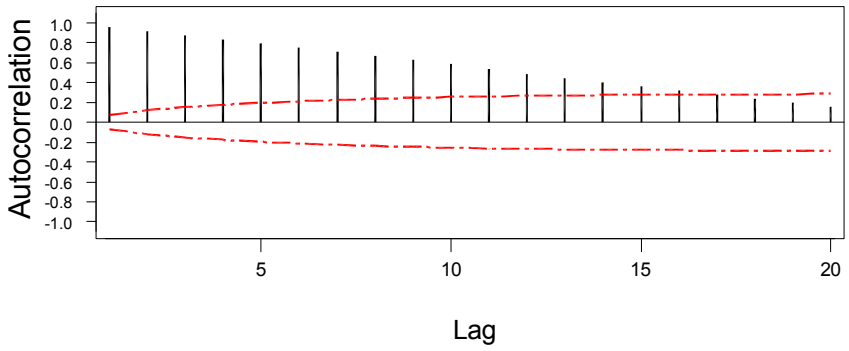


Figure A5.8. Autocorrelation plot of arsenic residuals

From Figure A5.9 the only significant coefficient is at lag 1. The statistician would recognise this as a key feature of a first-order autoregressive model — called AR(1). This simply means that the value of the variable at time (t + 1) is dependent on the value of the variable in the immediately preceding period; that is, at time t. One way of checking this assumption is to 'difference' the series in which a new observation is formed by differencing successive pairs of the original series. Figure A5.10 is a plot of the differenced arsenic residuals, which shows that all structure has effectively been removed and what remains is essentially 'white noise' — that is, random error (albeit with a non-constant variance).

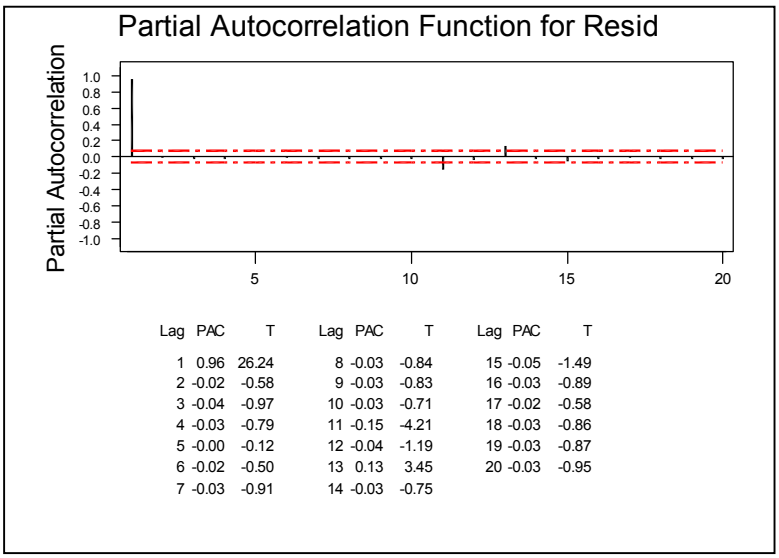


Figure A5.9. Partial autocorrelation function for arsenic residuals

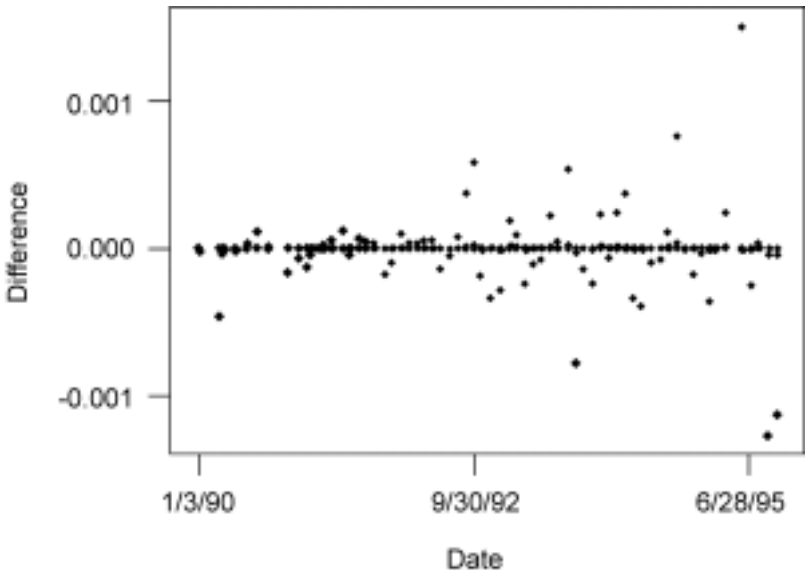


Figure A5.10. Time series plot of first differences of arsenic residuals

A5.1.5. Statistical Inference

The key features of inferential statistics are summarised in Figure A5.11. A population is the largest entity about which some assessment is required (e.g. a lake, a river, a beach, a reservoir, an ocean). Numerical quantities that describe some key feature of the population are referred to as parameters and are designated by Greek symbols. Three common population parameters are illustrated in Figure A5.11.

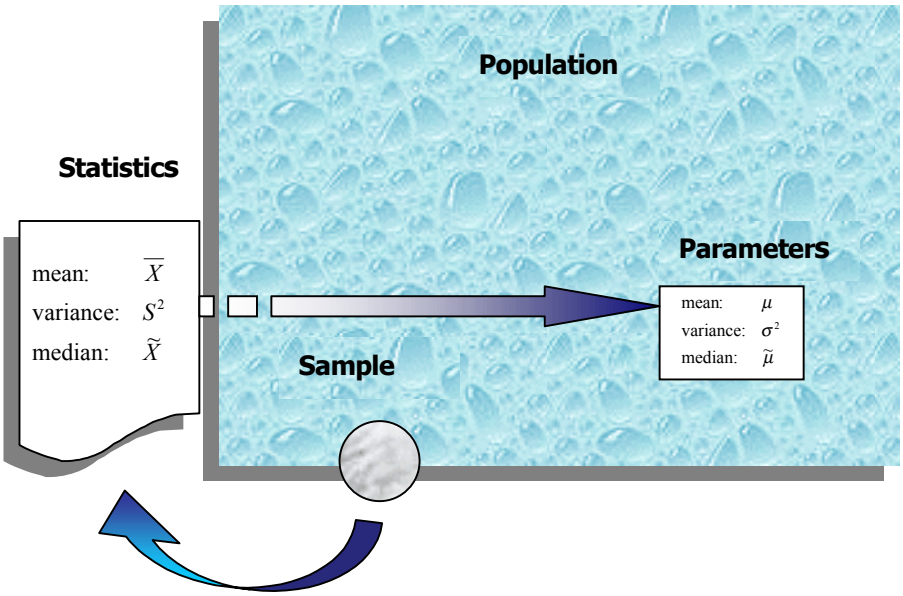


Figure A5.11. Relationship between population parameters and sample statistics

It is usually physically impossible and/or undesirable to sample an entire population and therefore decisions are based on the information contained in a randomly selected sample from the population. The numerical quantities that are computed for the sample are referred to as sample statistics. Sample statistics (or functions of them) are used to make inference about the true (but unknown) population parameters. Methods of inferential statistics fall into two main categories: estimation and hypothesis testing. There is commonality between these two activities although the emphasis is slightly different. Estimation is concerned with assigning a value (or range of values) to the true parameter while in hypothesis testing the aim is to make a judgement about the plausibility of a claimed parameter value. Whereas the reality of sampling is as depicted in Figure A5.11 (that is, a single sample from a larger population of values), methods of statistical inference are established on the notion of repeated sampling from the population. By examining the variation in the values assumed by a sample statistic under repeated sampling, statisticians are able to devise tests that make judgements about the likelihood of a particular result being attributable to chance variation or to some non-sampling induced effect (such as an incorrectly specified hypothesis).

A5.1.5.1. Interval Estimation

The concept of confidence intervals is introduced in section 6.4.1 in Chapter 6. To illustrate the underlying concepts, consider the hypothetical distribution of nitrate concentrations for all water bodies in some region of the country as depicted by the log-normal distribution of Figure A5.12a. The true mean concentration is 14.80 µg/L and the true standard deviation is 10.03 µg/L. The histogram in Figure A5.12(b) represents the distribution of averages obtained by repeatedly taking samples of size n = 20 from the distribution in Figure A5.12b. Some important points emerge:

- the population distribution (Figure A5.12a) is decidedly non-normal, yet the histogram of sample means (Figure A5.12b) shows a high degree of normality;
- the 'centre' of the histogram in Figure A5.12b is 14.713, which is very close to the true mean of the parent population, in Figure A5.12a, of 14.80;
- the standard deviation of the histogram of sample means is 2.113, which is less than the standard deviation of the parent population (10.03);
- about 95% of the sample averages in Figure A5.12b are contained in the interval 11 to 19.

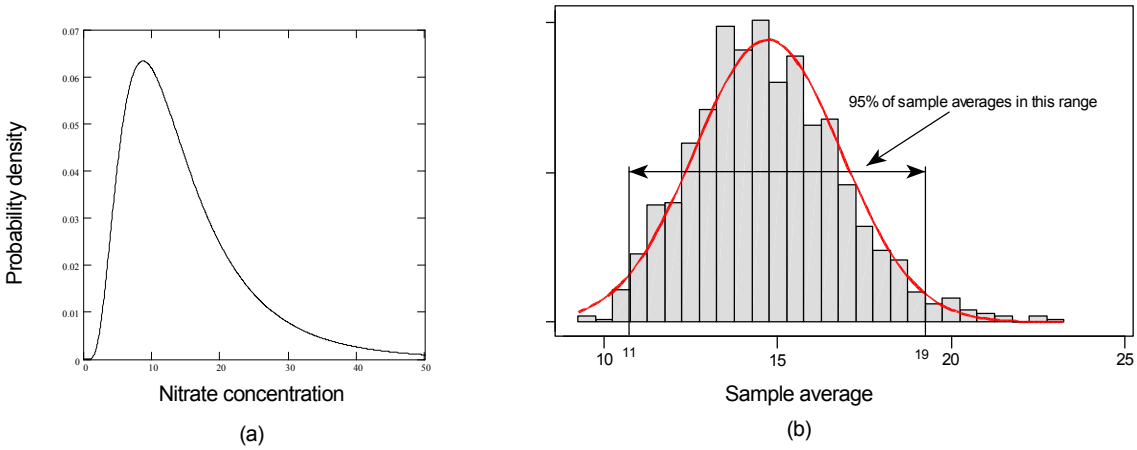


Figure A5.12. (a) Hypothetical distribution of nitrate concentrations for some population; (b) histogram of sample averages based on samples of size n = 20

These four observations are encapsulated in the Central Limit Theorem (CLT) in statistics. The CLT states that:

- the distribution of sample means is normal or approximately so, even if the parent population is non-normal;
- if the mean and standard deviation for the population are μ and σ respectively, then the mean of the distribution of sample averages is also μ, while the standard deviation is $\sigma/\sqrt{n}$ where n is the size of the sample from which the mean was computed. Note, for the example above, σ = 10.032 and so $\sigma/\sqrt{n} = 10.032/\sqrt{20} = 2.24$ which is in close agreement with the standard deviation of the histogram (2.113).

Finally, it is known from elementary statistics that approximately 95% of all observations lie within two standard deviations of the mean. Thus, we would expect that about 95% of our sample means in Figure A5.12b would lie in the interval $\mu \pm 2.0\,(\sigma/\sqrt{n})$; that is, between 10.31 and 19.29 (this is in good agreement with the observed range of 11 to 19). The interval 10.31 to 19.29 is called a 95% confidence interval for the true mean, μ.

To alter our level of confidence, the width of the confidence interval must change — that is the multiplier of 2.0 above needs to change. The appropriate multiplier is found by reference to tables of the standard or unit normal distribution. The general formula for a (1–α)100% confidence interval for a population mean, μ is

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where (1–α) is the level of significance (as a decimal between 0 and 1) and $z_{\alpha/2}$ is a value from the standard normal distribution (i.e. zero mean and unit variance), so that the area under this normal curve to the right of $z_{\alpha/2}$ is α/2. Table A5.3 gives some common levels of confidence and associated z multipliers.

Table A5.3. Critical z scores for selected levels of confidence

Level of confidence	**90%**	**95%**	**99%**	**99.9%**
z-value	1.645	1.96	2.578	3.291

Confidence intervals provide a good way of comparing observations of a water quality variable to a guideline value, because overlap between the confidence limits and the guideline indicate a lack of evidence that the guideline has not been exceeded.

There is a practical difficulty with the equation given above: namely, it requires knowledge of the true population standard deviation, σ. It is invariably the case that σ will be unknown and must therefore be estimated from the sample. It is tempting to simply replace σ by its sample estimate *s* in the equation above, but this cannot be done without additional modification. Without going into the theory, we need to replace the *z*-value with a value from the *t* distribution. The *t* distribution is very similar to the normal distribution, but it exhibits a degree of spread that is dependent on the sample size *n*. This dependence is expressed by a quantity called the degrees of freedom (df), and for the *t* distribution df = $n - 1$. The effect of various degrees of freedom on the *t* distribution is illustrated in Figure A5.13.

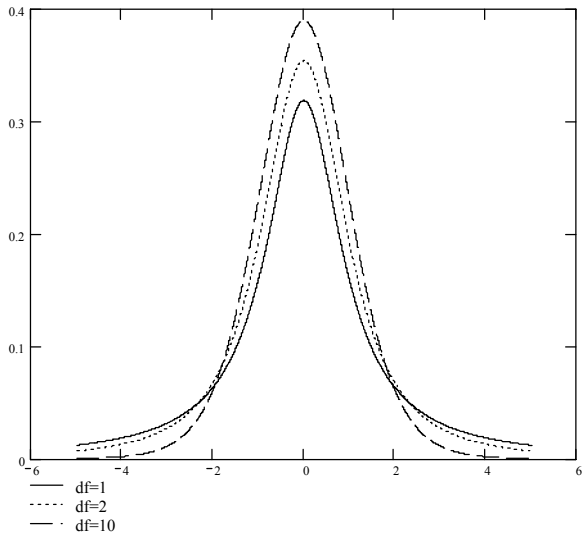


Figure A5.13. *t* distributions for df = 1, 2, and 10

Thus, whenever the sample size is 30 or less the following formula should be used in conjunction with the critical *t* values given in Table 6.6:

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

For sample sizes greater than 30, the *t* distribution is well approximated by the normal distribution. In such cases the normal distribution can be used to determine the 95% confidence limits.

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

Alternatively, a confidence interval for the population *median* may be used in place of the population mean. This is the preferred approach when the data exhibit a high degree of non-normality and mathematical transformation cannot increase the normality to an acceptable level.

The steps involved are listed in Table 6.9(a). The data are first arranged in ascending order. The two confidence limits are obtained as the C th observation from each end of the ordered data set. Determination of C depends on the sample size n. When n is 'large' (>20), the formula

$$C = (n - z_{\alpha/2}\sqrt{n})/2$$

is used. Note that C is rounded up to the next highest integer. For small values of n (≤ 20), C is obtained from Table A5.4.

Table A5.4. Values for C for samples of size $n \leq 20$

Sample size	Level of confidence			Sample size	Level of confidence		
n	**0.90**	**0.95**	**0.99**	***n***	**0.90**	**0.95**	**0.99**
4	1	–	–	13	4	3	2
5	1	–	–	14	5	4	3
6	2	1	–	15	5	4	3
7	2	1	–	16	5	4	3
8	3	2	1	17	6	5	4
9	3	2	1	18	6	5	4
10	3	2	1	19	6	4	4
11	3	2	1	20	7	6	5
12	4	3	2				

Example

Determine an approximate 95% confidence interval for the true population median phosphorus level (µg/L) using the following sample data:

4.86, 6.08, 4.83, 4.66, 4.98, 6.16, 6.08, 5.86, 3.95.

The ordered data set is:

3.95, 4.66, 4.83, 4.86, 4.98, 5.86, 6.08, 6.08, 6.16.

Using Table A5.4, $C = 2$ for $n = 9$ at 95% confidence. Thus the approximate 95% confidence limits are the 2nd observations from each end of the ordered data set; that is {4.66, 6.08}.

By way of comparison, a confidence interval for the *mean*, using the formula in section 6.4.1 and a critical t value of 2.306 from Table 6.6 for 8 degrees of freedom ($n - 1$), is

$$5.273 \pm 2.306\left(\frac{0.7916}{\sqrt{9}}\right), \qquad \text{i.e. } \{4.66, 5.88\}.$$

A5.1.5.2. Hypothesis Testing

Several different systems of statistical inference have evolved, and none of them is entirely satisfactory. The major difference between competing systems stems from their attitude to probability. For Bayesian statisticians, probability represents degree of belief in a proposition. This Bayesian view of probability is essentially subjective; probability represents an individual's belief about a system rather than an objective feature of the system. For frequentist or classical statisticians, probability represents the hypothetically limiting relative frequency of an event in a hypothetically infinite repetition of some chance system. For the frequentist statistician, probability represents an objective attribute of the chance system.

The logic of hypothesis testing can be illustrated with a simple example. Suppose the operator of a waste water treatment plant claims that recent upgrades in infrastructure will improve annual compliance with respect to some water quality parameter from the present 70%. The null hypothesis is a minimalist statement that effectively assumes nothing or assumes the status quo. In this example, the null hypothesis (H_0) states that the upgrades have made no difference to the level of compliance. On the basis of sample information, the operator hopes to obtain results that are incontrovertible in their support for the alternative hypothesis — that the true compliance has indeed increased. The parameter to be tested in this example is θ, the probability of compliance on any particular occasion. The situation is reflected by the following pair of hypotheses:

$$H_0: \ \theta = 0.7, \qquad H_1: \ \theta > 0.7 \ .$$

Suppose monthly samples are collected and over the following year the regulator notes that the operator is in compliance on 9 of the 12 occasions, or 75% of the time. Although 75% certainly represents an improvement, the issue to be resolved is whether or not 75% compliance (or better) is likely to occur due to chance even when no improvements have been made. In fact, it can be established that there is about a 50% chance of this happening. This is hardly compelling evidence, and so the decision would be to accept the null hypothesis of no improvement[2]. While this example may seem rather trite, it nevertheless succinctly illustrates a number of key concepts associated with statistical hypothesis testing. First, the analyst is faced with making a binary choice — accept the null hypothesis or reject it[3]. Secondly, this decision is based on incomplete information and thus the decision-making process is subject to error. The types of error possible are identified in Table 6.7 in Chapter 6.

Since the true state of nature is never known, the decision-making process is based on an assessment of probabilities concerning the outcomes identified in Table 6.7. The probability of committing a Type I error is called the level of significance[4] and is denoted by the Greek symbol α. The monitoring team has complete control over α and must specify it in advance of any data analysis. A liberal statistical test will result from assigning a relatively high value for α (since the test procedure will tend to lead to a 'reject' decision more often) whereas setting α to be very small ensures a conservative test. The probability of committing a Type II error (designated by the Greek symbol β) will generally remain unknown because this probability will, in part, depend on the degree to which the null hypothesis has been violated.

A particularly important concept related to β is the complement 1 – β, which is referred to as the *power* of the statistical test. The power of a statistical test is a numerical measure of its ability to correctly reject a false null hypothesis.

In developing hypothesis tests, statisticians first fix the Type I error rate at some small value, typically 0.05 or 0.01. The 'best' decision-rule is the one for which the Type II error rate is as small as possible for this fixed Type I error rate. This treatment of Type I and Type II error rates is asymmetric, in this formulation it is more important to reduce Type I errors.

A5.1.6. The Two-sample *t*-test (Independent Samples)

Often a monitoring team needs to compare two populations. In Worked Example 2 (page A5-29), the analyst wants to know if there is a difference in cadmium concentration between male and female oysters. The two samples have different sample means and sample variances, although this is to be

[2] The probability of getting at least 11 compliances, when in fact the null hypothesis is true, is about 9%; this is still generally not regarded as small enough to assert the veracity of the alternative hypothesis 'beyond reasonable doubt'. Clearly this 'experimental design' appears to disadvantage the operator who would need to demonstrate 12 out 12 compliances in order to 'prove' greater than 70% compliance.

[3] This binary decision making process has been criticised for the narrow focus it forces the analyst to adopt.

[4] The level of significance is also referred to as the *size* of the statistical test.

expected even if the samples have been drawn from the same population. To determine whether or not the observed difference between sample means has occurred simply by chance or whether it reflects a true difference between sexes, a Student's *t*-test can be performed. If the probability (*p*) that the two sample means are from the same population is found to be small (say $p<0.05$) it can be concluded that the two population means are probably different. If the probability is large (say $p>0.05$) the conclusion will be that the difference between means is more likely to be a result of sampling variation.

Before a Student's *t*-test is conducted, the assumption of homogeneity of variances needs to be checked. A number of tests are available for this purpose. If only two variances are being compared, an *F*-test can be used. A generalisation of this test is Hartley's F_{max} test based on the ratio of the largest sample variance to the smallest sample variance; see Ott (1984) for details and a table of critical values. If the variances are found to be not significantly different from one another (see Worked Example 7, page A5-36) a Student's *t*-test may be used to test the means. If the variances are significantly different, a Student's *t*-test would be inappropriate. In this situation Welch's *t*-test, which does not assume homogeneity of variances, could be applied instead.

A5.1.7. The Two-sample *t* -test (Dependent Samples)

The preceding *t*-test required samples to be independent. In cases in which there is a pairwise dependency between samples, the paired *t*-test is preferred: for example, when comparing growth rates of an aquatic plant before and after the administration of a substrate nutrient. A pair of growth measurements is taken from each plant, one measurement before and one after the administration of the nutrient. With paired samples, the procedure is to calculate the difference between members of the pairs, and then to test whether the mean of these differences (as opposed to the difference in means used by the independent samples *t*-test) is significantly different from zero. The paired *t*-test is a more powerful analysis than the independent samples *t*-test because each pair acts as its own 'control', thereby reducing variation due to factors not being investigated. An example of a paired *t*-test is illustrated in Worked Example 3 (page A5-31).

A5.1.8. Analysis of Variance

Analysis of Variance (ANOVA) represents a logical extension of the two-sample *t*-test. ANOVA is a well-established parametric[5] technique for testing the hypothesis of simultaneous equality of a collection of population means. Its major uses in the analysis of water quality data include assessment of the significance of differences in a measurement among water bodies, among different localities in the same water body, or among samples taken from one location at different times.

ANOVA techniques assume that the data in each 'treatment' group are drawn from normally distributed populations, with common variance. The ANOVA procedure is an ingenious device for making inference about the simultaneous equality of the group means by examining components of variance — hence the name. It is possible to obtain two separate estimates of the common variance, σ^2. One estimate, referred to as the within-groups variance, is based on a pooling or 'averaging' of the individual group variances. The second estimate, known as the between-groups variance, uses the fact that the variation between the group means should be roughly σ^2/n. Therefore, if we estimate the variation between the group means and multiply that value by *n*, the size of each group, we obtain an estimate of σ^2. When the null hypothesis is true, these two estimates should be approximately equal, or in other words the ratio of between-group variance to within-group variance should be approximately one. However, when the null hypothesis is false, it can be shown that the between-group estimate is inflated and thus the ratio will return results greater than one. This ratio follows an

[5] In this document the term parametric is used to distinguish statistical methods whose application assumes the sample data follow some prescribed probability model. Statistical procedures that relax this assumption are referred to as nonparametric or 'distribution-free' methods.

F-distribution, and so reference to statistical tables allow us to decide if a particular value of the ratio is too large to ascribe to chance alone when the null hypothesis is true.

Worked Examples 4, 5 and 6 (pages A5-32 to A5-35), are examples of single factor ANOVAs. In Worked Example 4 the analyst wishes to determine if there was a difference in the abundance of mayfly larvae among four shorelines (north, south, east, west). Shore direction is the 'factor', with four discrete 'factor classes', and mayfly abundance is the 'response variable'.

A5.1.8.1. Multiple Comparison Procedures

The rejection of the null hypothesis in an analysis of variance is a rather empty achievement — it enables one to assert with some degree of confidence only that there are differences between the group means. In the example above, the obvious next question is which sites or times are different from which others. To help answer this and similar questions, the monitoring team may employ a multiple comparison technique to help discover underlying causes for the rejection of the null hypothesis.

There are a number of multiple comparison procedures available and they tend to differ with respect to the type of overall error to be controlled for. Procedures for comparing all means with all others include Fisher's least significant difference (LSD), Tukey's multiple comparison test procedure and the Student–Newman–Kuels (SNK) method. Dunnett's test is appropriate where there is a single control site (say upstream of a discharge) and a number of treatment sites (say various distances downstream) with which this is to be compared (see Worked Example 5).

A5.1.8.2. Fixed Versus Random Effects

The above examples are known as fixed single-factor ANOVAs because the factor classes were deliberately chosen. They are fixed in the sense that if the experiment were to be repeated again, the same factor classes would be chosen. This design is the most commonly encountered type in water studies. There is also a random single-factor ANOVA in which the factor classes are randomly chosen from a pool of possible choices. For example, consider the problem of assessing the analytical capabilities of laboratories in measuring chlorophyll-*a*. If there are too many laboratories for all of them to be included, and the monitoring team is not interested in any particular laboratory, a random selection from the pool of possibilities can be made.

A5.1.8.3. Replication and Power

An important consideration in the design of any study is the level of replication to be used. The number of replicates required should be thought about at the design stage of the study. Important factors to consider are: the magnitude of the smallest differences that need to be detected; an acceptable probability that such a difference will not be detected; a notional indication of how variable replicates at a single site and time are likely to be (as indicated by a pilot study); and the significance level of the test to be used. All these considerations are central to a formal power analysis (discussed later in this appendix, page A5-20).

A5.1.8.4. Use of Controls

Another design consideration is the identification of controls. In the context of impact studies (e.g. the effects of an industrial development), a control may be temporal or spatial.

Temporal controls involve the collection of data before an industrial development (say) that may have caused an effect, and form the basis of comparison with post-impact data. The limitation of temporal controls is that there may have been a temporal trend regardless of the industrial development. Attributing a difference between pre- and post-development data to the effect of the development therefore assumes no independent temporal trend. A case must be made in support of this assumption in studies involving only a temporal control.

Spatial controls involve collection of data at a distance from the area likely to have been affected by the development, say upstream from a discharge site, and they form the basis of comparison with the affected sites. The limitation of spatial controls is that the control area and the area subject to potential impact may have been different in the first place. The upstream site may have been different from the downstream sites before the industry was developed and before waste was discharged into the stream. Attributing a difference to the effect of the development assumes there was no prior difference between the control and potentially affected areas. A case must be made in support of this assumption in studies involving only a spatial control.

Other designs have both temporal and spatial controls. Demonstration, before the development, that there is no difference between the area to be developed and the spatial control imparts a certain degree of confidence that any later difference between the affected area and the spatial control area is a consequence of the development. Similarly, demonstration that there is no temporal trend in the control site(s) provides confidence in an interpretation that a difference between pre- and post-development is a direct effect of the development. It is necessary, however, to ensure no spatial–temporal interaction. That is, the monitoring team must be confident that in the absence of the industrial development, the sites that are to be developed would not behave any differently through time compared to the control sites.

A5.1.8.5. Factorial Analysis of Variance

Factorial ANOVA involves the simultaneous analysis of the effects of more than one factor on the measurement variable. Its simplest form, the two-factor design, is more powerful than two separate single factor designs because the effect of one variable can be partitioned out before the effect of the other is tested. For example, in Worked Example 8 (page A5-37), a two-factor ANOVA shows that there was a significant difference in the body mass of *Antechinus stuartii* between the years 1977 and 1978. However, if analysed as a single-factor ANOVA, the difference between years was not significant (Table A5.5). This is because the variation within years was inflated by differences in body mass between sexes (which is demonstrated by the significant result between sexes in the two-factor ANOVA). In the two-factor ANOVA, variability between sexes was partitioned out before years were compared (compare the within-groups sums of squares between the two analyses: the difference is due to sex, 191+333=524). This resulted in a smaller mean square error term and hence a more powerful analysis. Although the researcher may not have been particularly interested in difference between sexes, the inclusion of this factor allowed for a more sensitive test of the temporal effect.

Table A5.5. Results of a single-factor ANOVA on the body mass of *Antechinus stuartii* between the years 1977 and 1978. The data are the same as those used in Worked Example 7 (p. A5-36)

Source	DF	SS	MS	*F*	Pr > *F*
Among groups	1	162.56	162.56	4.34	0.056
Within groups	14	524.38	37.46		
Total	15	686.94			

The two-factor analysis also enables a test for interaction effects. In the above example, body mass may differ between years, but the magnitude of the difference, or, in extreme cases, the direction of the difference, may depend on the sex. In this case the two factors are said to interact.

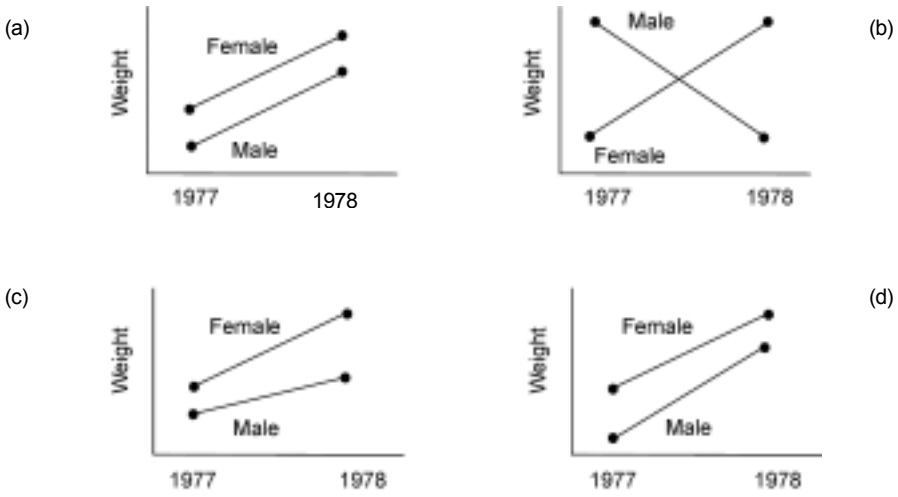


Figure A5.14. Various forms of interactions between year and sex on body mass: (a) represents no interaction; (b) represents strong antagonistic interaction; (c) represents synergistic interaction; (d) represents unimportant interaction

A significant interaction can provide difficulties in the interpretation of the main effects because they must be described in terms of the combined effects of both factors. Figure A5.14b shows a case of an extreme interaction. Differences in mass between sexes are dependent on the year under investigation. Males were heavier in 1977 while females were heavier in 1978. Conversely, differences between years depend on which sex is being investigated. The 1977 samples had heavier males than the 1978 samples, but lighter females. Therefore, in the presence of significant interaction, tests of main effects cannot be made independently of each other. Some interactions (Figure A5.14c) can be corrected by transformation, enabling the main effects to be tested separately. However, a transformation would be of little benefit for interactions such as Figure A5.14b. Some interactions can be considered unimportant (Figure A5.14d) because the effect of one factor on the other is relatively small. In such cases the interaction can be ignored and the main effects tested separately.

A basic strategy in interpreting the results from a factorial ANOVA is to follow these six steps (Neter et al. 1996):

1. examine whether there is an interaction between factors;
2. if there is no interaction, test main effects;
3. if factors interact, determine whether the interaction is important;
4. if interaction is unimportant, proceed as in step 2;
5. if interaction is important, attempt transformation to make it unimportant and return to step 2;
6. if interaction is still important, analyse the two factor effects jointly in terms of the treatment means (e.g. compare body mass between male and female *Antechinus* sp. separately for the years 1977 and 1978, rather than combining the years). See Neter et al. (1996) for more information.

In two- and higher-order factor ANOVAs there are three types of model: fixed factor (all factors are fixed; random factor (all factors are random); and mixed factor (some factors are fixed and some are random). Another important distinction between models is the calculation of the *F* statistic. In a fixed model, the *F* statistic is calculated by dividing the mean square of each factor by the within (error) mean square. In a random model, the *F* statistic is determined by dividing the mean square of both random factors by the interaction mean square, while the interaction is divided by the mean square within. The mixed model is different again. For the calculation of the *F* statistic for each model, see Table A5.6.

Table A5.6. Computation of the *F* statistic for tests of significance in a two-factor ANOVA with replication (Zar 1984); MS is mean square

Effect	Model I (factors A and B fixed)	Model II (factors A and B random)	Model III (factor A fixed, factor B random)
A	MS_A/MS_{within}	MS_A/MS_{AB}	MS_A/MS_{AB}
B	MS_B/MS_{within}	MS_B/MS_{AB}	MS_B/MS_{within}
A B interaction	MS_{AB}/MS_{within}	$MS_{AB}/MS_{wiithin}$	MS_{AB}/MS_{within}

ANOVA allows for the assessment of multi-factor models. Theoretically the number of factors possible is unlimited, but models with five or more factors are rare because of the large number of experimental units that are required and the difficulty in interpreting higher-order interactions.

A5.1.8.6. Nested Analysis of Variance

In the factorial designs discussed above, the factors are crossed, i.e. all levels of one factor occur within all levels of the other factor(s). Both male and female *Antechinus* sp. are included in each year (1977 and 1978). However some experimental designs are not crossed because one factor occurs at different levels in combination with another factor. This is known as a nested design with one or more of the treatments nested within another factor. An example of a nested design can be found in Worked Example 6 (page A5-35), in which fluoride concentration is the dependent variable, and the factor 'sample' is nested within the second factor 'location'. There are three samples within each location, but these samples are unrelated. In this design, the nested factor (sample) is random and the primary factor (location) is fixed. In most cases the nested factor has no intrinsic interest — it is only included to account for some within-group variability. In the example in Worked Example 6, variability due to measurement error has been accounted for by the inclusion of the nested factor (sample). This increases the test's sensitivity to significant effects in the primary factor (location).

A5.1.8.7. Analysis of Covariance

Another important analysis in water quality studies is analysis of covariance (ANCOVA). Suppose a monitoring team wishes to compare the zinc concentration in a species of fish between two lakes (Worked Example 9, page A5-38). A non-significant result is obtained from a single factor ANOVA or *t*-test analysis (Table A5.7). This is because there is a strong negative relationship between zinc concentration and body mass in fish (see graph in Worked Example 9). Therefore, as a large range of fish sizes has been sampled, there is increased variability of zinc concentration in fish within lakes.

Table A5.7. Results of a single factor ANOVA on the zinc concentration in fish, between Lake Arthur and Lake Bull. Data are the same as those used in Worked Example 9.

Source	DF	SS	MS	*F*	*p*
Among	1	2408.33	2408.33	4.38	0.063
Within	10	5493.33	549.33		
Total	11	7901.67			

One way to account for this would be to impose a 'physical control' by using fish of the same size, thereby eliminating any 'size-induced' variation. However this could be time consuming as well as wasteful of data. An alternative strategy is to use a statistical adjustment for the effects of fish size. ANCOVA achieves this by testing for differences among group means after adjusting for group differences in the independent variable (known as the covariate). That is, it compares zinc concentrations in fish between lakes after adjusting for differences in the mass of fish. From Worked

Example 9, it can be seen that there is a significant difference in zinc concentration in fish between Lake Arthur and Lake Bull. Therefore, a large proportion of variability within each lake has been accounted for by the relationship between zinc concentration and body size (compare the within mean square between the two analyses: 11.18 for the ANCOVA and 549.33 for the ANOVA).

Where there are more than two treatments, multiple comparison tests can also be applied to determine which treatment differs from which, although it is important that treatment means are adjusted for the effect of the covariate first.

It should be noted that ANCOVA is different from a two-factor ANOVA, as the covariate (in this case fish mass) is a continuous variable, not a factor. In a two-factor ANOVA both independent variables are discrete. More complex models are possible. For more information on ANCOVA see Sokal and Rohlf (1995) or Neter et al. (1996).

A5.1.9 Generalised Linear Models

Although extremely powerful when attendant assumptions are satisfied, the classical methods of inference presented so far have limitations that restrict their more widespread applicability. For example, the use of ANOVA models to make inferences about the abundances of a rare species is likely to yield misleading results because of the inappropriate statistical model conferred by the ANOVA assumptions. In this instance, the response variable (observed counts) is assumed to follow a normal distribution. The use of a continuous probability model to describe an inherently discrete process may not be an issue if the mean of the counts is sufficiently large for the central limit theorem to apply. However, by definition, our rare species will be observed in very low numbers — perhaps with a mean of five or less and the errors introduced by the normal approximation to the underlying discrete distribution may be substantial. No amount of data transformation is likely to remedy the situation and a more appropriate statistical model should be sought rather than trying to coerce our data into a statistical straightjacket that does not fit. Another limitation of classical methods is the difficulty in modelling non-linear relationships between the response variable and the mean. It was these and other considerations that motivated Nelder and Wedderburn in 1972 to develop a more general class of statistical models known as Generalised Linear Models. There is a subtle distinction to be made here between this new class of models and the classical methods of inference that fall under the umbrella of *general* linear models. The generalisation brought about by Nelder and Wedderburn's development is a consequence of the following features:

- explicit handling of non-normal error distributions;
- easy accommodation of non-linear relationships between the response variable and the mean;
- ability to model both qualitative and quantitative data;
- ability to model processes that have intrinsic mean–variance relationships.

While the generalised linear model framework has led to an enrichment of statistical modelling capabilities, the uptake of these tools has been somewhat disappointing. This may in part be attributed to the more advanced statistical concepts, the introduction of new terminology, and a slightly different approach to analysis and interpretation. Further compounding this situation has been the lack of easy-to-use statistical software for fitting generalised linear models, although programs such as SAS® and S-PLUS® have this capacity.

The underlying theory of generalised linear models is beyond the scope of this document and requires solid foundations in linear algebra and calculus (see also Worked Example 10, page A5-40). The original paper by Nelder and Wedderburn is unlikely to be of benefit to readers of this document. More accessible accounts of generalised linear models and their applications may be found in McCullagh and Nelder (1983) and the text by Dobson (1990).

A5.1.10. Power Analysis and Sample Size Determination[6]

The main purpose of power calculations is to assist in the planning and design of monitoring programs. Before large resources are committed to monitoring, the monitoring team would like to be confident that the program is capable of detecting an effect that is considered large enough to be important, with a reasonably high probability. If a monitoring program has low power, it means that even effects large enough to be of interest are unlikely to produce statistically significant results. In this case, the analyst should consider the possibility of increasing the power (usually by using more replicates), or perhaps question whether the monitoring should be undertaken at all.

The power of a significance test is a concept rooted in the Neyman–Pearson view of hypothesis testing. Power is simply the complement of the Type II error rate: that is, power is the probability of *not* making a Type II error (i.e. of not having a false sense of security), given that the null hypothesis is false. A test with high power (at some specified effect size) is therefore very likely to detect effects of the given size.

In deciding on an appropriate level of replication, account must be taken of both Type I and Type II errors. Neyman and Pearson argued that the costs of Type I and Type II errors could not be defined formally, and that the decision procedure should be based on an informal balancing of Type I and Type II error rates. A Type I error rate of 5% has become conventional, and a Type II error rate of 20% (power 80%) is often considered acceptable. Of course, both of these error rates should be determined in advance and set in the context of the problem at hand. No statistical textbook explains how this ideal state is to be achieved.

Power calculations have also been used to assist in the interpretation of results that are not significantly different. Here their use is retrospective ('how do I make sense of the monitoring I performed?') rather than prospective ('how do I undertake monitoring that will be useful?').

There seems to be some uncertainty about how retrospective power calculations should be performed. Some practitioners perform retrospective power calculations at the effect sizes observed in the experiment. Arguably, there is little merit in this approach. Since retrospective power calculations tend to be computed only when the null hypothesis is accepted, the calculated power is likely to be very low. Indeed for a test statistic that has a symmetric distribution under the alternative hypothesis, the power calculated in this way will always be less than 50%. In the extreme case of no observed effect, the power will be equal to the significance level. What is important is the power of the experiment to detect results that are believed to be important. It is recommended that power calculations be performed for a range of effect sizes, spanning the region in which effects are large enough to be important.

The following section illustrates some of the basic concepts behind power and sample size calculations. Given the complexity of the calculations involved, no attempt is made to provide a comprehensive catalogue of formulae for various statistical test procedures — these are best left to reliable software tools. The calculations used in this section have been performed using CSIRO's free software package *PowerPlant*®. A copy of the software and documentation may be downloaded from the following site: ftp://ftp.per.its.csiro.au/csiro-wa/biometrics. A list of other power and sample analysis software tools can be found at: http://www.insp.mx/dinf/stat_list.html or http://www.forestry.ubc.ca/conservation/power/index.html .

A comprehensive review of a number of these utilities can be found in Thomas and Krebs (1997).

A5.1.10.1. Basic Concepts

Consider the following hypothesis-testing situation: H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$,

[6] Adapted from Statistical Power Analysis – Course Notes, CSIRO Mathematical and Information Sciences www.cmis.csiro.au/envir/Training/StatPowerAnal.htm .

where μ is the true but unknown mean concentration of some water quality parameter, and μ_0 is the hypothesised value. A graphical representation of the distribution of the parameter of interest under each of these hypotheses is shown in Figure A5.15 (note it is necessary to assume normality and a common variance σ^2).

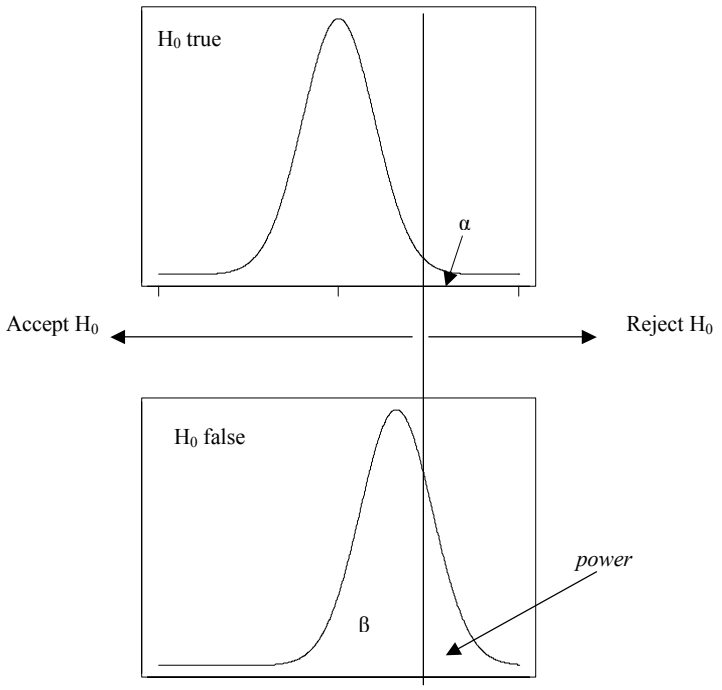


Figure A5.15. Level of significance (α), Type II error (β), and power (1 – β)

When the null hypothesis is true (top curve in Figure A5.15), the probability of incorrectly rejecting H_0 (α) is represented by the area to the right of the decision rule (the vertical line in Figure A5.15). Conversely, when the null hypothesis is false (bottom curve in Figure A5.15), the probability of incorrectly accepting H_0 (β) is represented by the area to the left of the decision rule (the vertical line in Figure A5.15). The power (1 – β) is the probability of correctly rejecting a false H_0 and this is depicted as the area to the right of the decision rule under the bottom curve of Figure A5.15. This simple example serves to highlight the fact that the power depends on the degree to which the bottom curve of Figure A5.15 has been displaced relative to the top curve (i.e. the bigger the 'effect size', the greater the power of the statistical test). Power is also affected by sample size, level of significance, and variance of data.

Example

Consider the situation in which the concentration of some analyte is to be compared at five locations using a one-way ANOVA design. The null and alternative hypotheses are:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ versus H_1: at least two means different.

The monitoring team statistician has decided on using $n = 4$ samples at each of the five sites. From past experience she knows that the standard deviation is approximately 0.14 µg/L. It is important that the experimental design has sufficient power to detect a minimum difference of 0.02 µg/L using a 0.05 level of significance. The dialogue box from the *PowerPlant*® software is shown in Figure A5.16.

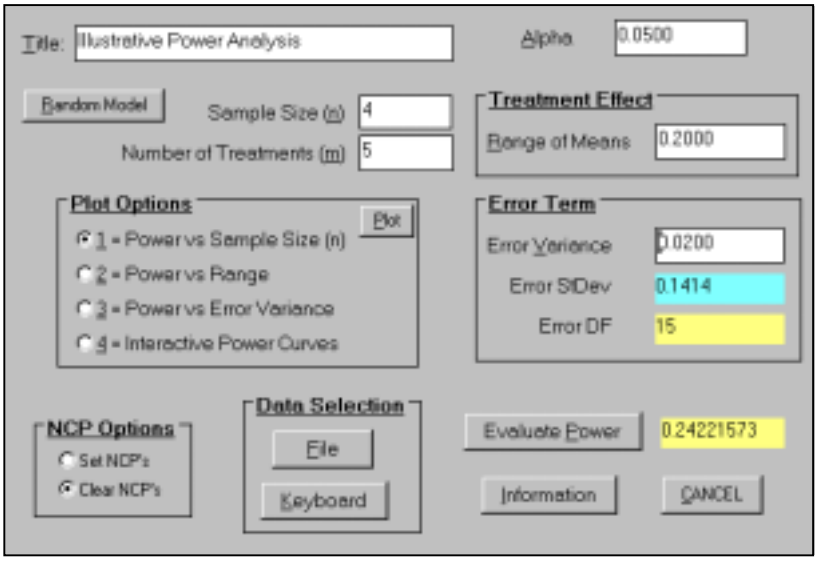


Figure A5.16. *PowerPlant*® dialogue box for power and sample size calculations

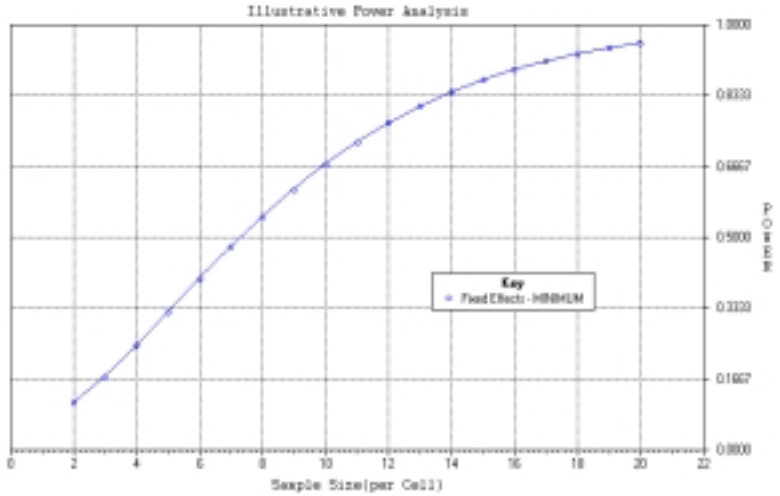


Figure A5.17. Power curve from *PowerPlant*® software

It can be seen from Figure A5.16 that the power of the current experimental design is a little under 25%. This would be regarded as too low to be of benefit and the decision must be made to either abandon the experiment or increase resources (more samples). To help address the latter, it is useful to obtain a plot of the power curve as a function of sample size. This is simply a matter of selecting the appropriate plot option in the *PowerPlant*® dialog box and pressing the 'Plot' button. The result is displayed in Figure A5.17. Figure A5.17 shows that approximately 13 samples at each site need to be taken in order to have a power of approximately 80%. (See also Worked Example 11, page A5-41.)

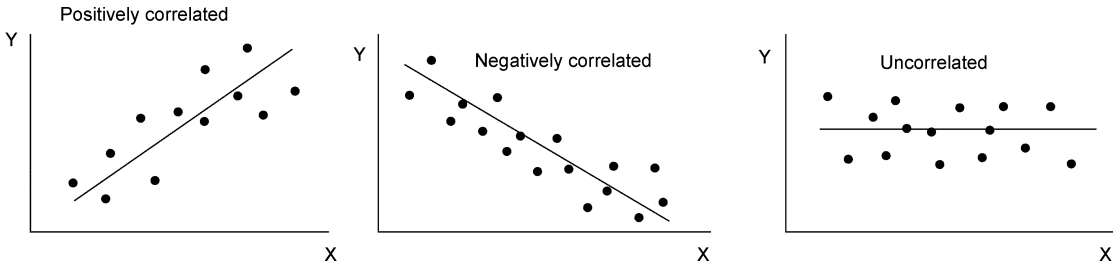


Figure A5.18. Examples of positively correlated, negatively correlated and uncorrelated data

A5.1.11. Correlation and Regression

Correlation analysis refers to the statistical methods associated with quantifying the (linear) relationship between two or more variables. Examples of the types of relationships observed are shown in Figure A5.18. Regression analysis refers to statistical procedures that attempt to *model* or describe (in the form of mathematical functions) that relationship. Both correlation and regression are discussed in section 6.5 in Chapter 6. Regression is illustrated in Worked Example 12, page A5-42.

For simple linear regression, the following model is assumed,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ ,$$

where β_0 is the true (but unknown) intercept, β_1 the true slope and ε_i the ith residual. The latter terms are introduced to account for the imperfect fit between Y and x. Furthermore, these residuals are assumed to follow a normal distribution having mean zero and variance σ_ε^2. The regression analysis usually commences (and unfortunately, all too often, stops) with the estimation of the parameters β_0 and β_1. The standard statistical procedure for accomplishing this is known as ordinary least squares (or OLS). Formulas for estimating β_0 and β_1 from sample data are provided below (a 'hat' over a β differentiates an estimate of a parameter value from the parameter value itself):

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

where

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right); \quad S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 .$$

Regression models are used in a variety of contexts. For example, a monitoring team might wish to develop an empirical model to predict chlorophyll-*a* concentrations from loads of phosphorus. Regression can also be used as an exploratory tool to investigate relationships between a response variable and other environmental variables with a view to formulating hypotheses for subsequent manipulation and experimentation to establish causation.

Aquatic systems are often too complex to be modelled by simple linear regression. For example, chlorophyll-*a* concentrations would be dependent on more than just the loads of phosphorus alone. Other variables such as flow rate, water temperature and nitrogen concentration may also be important in determining the concentration of chlorophyll-*a*. Therefore, a simple linear model is likely to be a poor predictor of variation in chlorophyll-*a* concentrations. In such cases a multiple regression model may be more appropriate. The multiple regression model takes the form

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \ .$$

The interpretation of the βs is that they represent the change in Y per unit change in the corresponding X, assuming all other Xs are held constant.

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, if a measurement taken at one time is above the value predicted from the statistical model under consideration (in this case a regression line), it is likely that the next value will also be above its predicted value. 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.

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

A5.1.11.1. Robust Regression

Robust regression is discussed in section 6.5.3 in Chapter 6. To illustrate some of the concepts, consider the metal data of Figure 6.2. Inspection of Figure 6.2 shows that although the relationship between lead and copper concentrations exhibits a high degree of linearity, there are two lead values which appear to be atypical. The Cu – Pb scatter plot is shown in Figure A5.19 together with the regression line estimated using OLS. The influence of the two aberrant lead values is most pronounced and while the fitted line is the 'best' in terms of the least-squares criterion, its predictive capability would be low. To overcome this difficulty we could remove the offending observations and re-fit the line. This may be appropriate in this instance since the source of the problem is clear. However, in some instances it may not be obvious which observations have high 'leverage' and so a robust method using all the data would be preferred. The robust regression analysis of the Cu – Pb data is shown in Figure A5.20. The resulting line provides a much more realistic fit to the data.

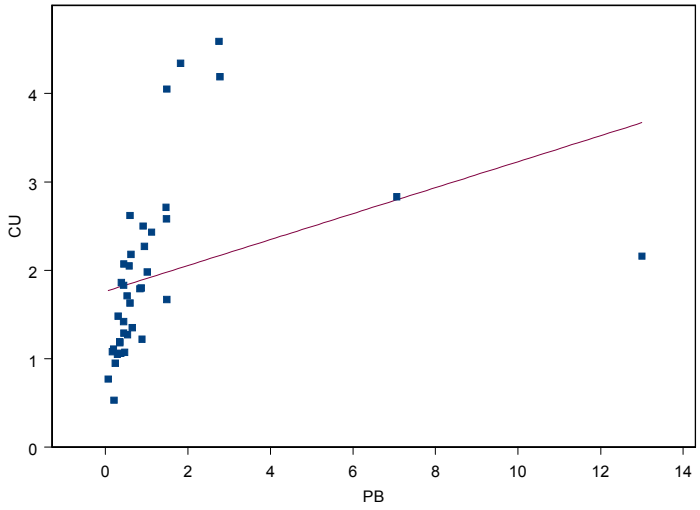


Figure A5.19. Scatterplot for copper and lead concentration data of Figure 6.2 with OLS regression line overlaid

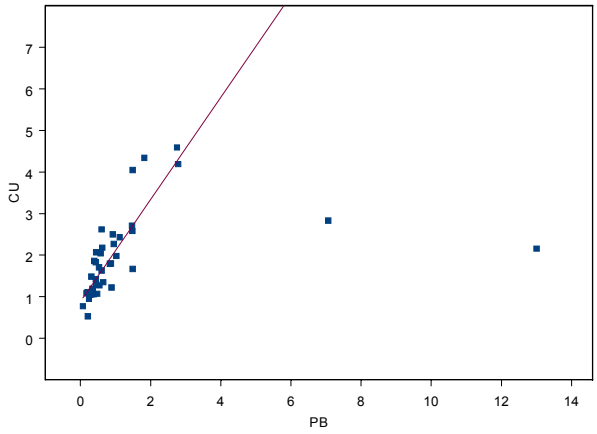


Figure A5.20. Bi-plot for copper and lead concentration data of Figure 6.2 with robust regression line overlaid

A5.1.12. Generalised Additive Models

In the same spirit as the generalised linear models described earlier, generalised additive models (GAMs) have been devised to increase the flexibility of statistical modelling. As noted by Hastie and Tibshirani (1990), 'we can now augment the linear model with new methods that assume less and therefore potentially discover more'. The GAMs represent an extension of conventional regression modelling techniques. Rather than imposing and estimating some predefined model, GAMs replace the usual linear function of an independent variable with 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.

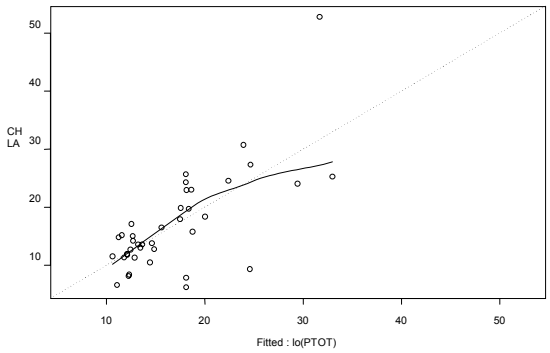


Figure A5.21. Plot of measured chlorophyll-*a* and fitted values using a smooth function of phosphorus

By way of example, consider modelling the relationship between chlorophyll-*a* and total phosphorus in a lake using a generalised additive model. Rather than deciding on some functional form in advance and then estimating the parameters of our model, we can use a GAM to help determine the nature of this relationship. Figure A5.21 shows the result after fitting chlorophyll to a smooth function of total P. The dashed line in the plot is used to gauge the degree of departure from a linear model. From Figure A5.21 we see the departure is quite small, although some non-linearity is evident at higher chlorophyll levels. In order to assess the adequacy of the fit a comprehensive statistical analysis of the resulting model is possible, but this requires an understanding of more advanced concepts. These are covered in the text by Hastie and Tibshirani (1990).

A5.1.13. Nonparametric Statistics

Parametric tests such as *t*-tests and ANOVA have several restrictive assumptions, such as homogeneity of variances and data that are normally distributed. Violations of these assumptions can result in seriously flawed decision-making. In situations where the assumptions of a parametric test cannot be met, non-parametric tests or 'distribution-free' counterparts may be more appropriate. Some of the more common parametric tests and their nonparametric alternatives are listed in Table A5.8. For more detailed discussion on nonparametric tests see Neave (1988).

Table A5.8. Common parametric tests and their nonparametric alternatives, and the advantages and disadvantages of nonparametric tests

Parametric test	Nonparametric test
Paired *t*-test	Wilcoxon matched pairs signed ranks test
Students *t*-test	Mann–Whitney U
	Kolmogorov–Smirnov two-sample test
One-way ANOVA	Kruskal–Wallis H test
Two-way ANOVA	Friedman two-way ANOVA for ranks
	Scheirer–Ray–Hare extension of the Kruskal–Wallis test
Linear regression	Kendall's robust line-fit method
Pearson's correlation *r*	Spearman's rank correlation
	Kendall's rank correlation

Advantages of nonparametric tests	Disadvantages of nonparametric tests
• Free of most distributional assumptions	• Generally less powerful than parametric tests
• Resistant to outliers (this may also be seen as a disadvantage)	• Do not always make most use of information contained in data
• Generally easy to perform	• There are no non-parametric alternatives to some of the more complex parametric analyses such as multiple regression and multi-factorial ANOVA
• Potentially more robust in the presence of non-detects	
• Handle nominal and ordinal data	

A5.1.14. Multidimensional Scaling

A detailed account of multidimensional scaling (MDS) is beyond the scope of this document and the interested reader is advised to consult one of the numerous texts on multivariate statistics, e.g. Clarke and Warwick (1994). This section provides a very brief overview of the method and its uses, and an illustrative example (see also section 6.6.3 in Chapter 6).

Multidimensional scaling is usually introduced by analogy with the following situation. Given a map showing the locations of major cities it is a straightforward task to construct a table of distances between the cities. However, the reverse problem is considerably more difficult. That is, given a table of inter-city distances, how is the map constructed? It is this latter problem that MDS attempts to resolve. The difficulty is compounded in the natural environment because our input data (the

'distance' matrix) is subject to considerable uncertainty, and, furthermore, the number of dimensions required to adequately represent the observed similarities or distances is not always evident. As noted in section 6.6.3 in Chapter 6, the computations required to undertake an MDS analysis require access to specialised statistical software.

An example of an MDS plot is shown in Figure A5.22. In this case researchers were interested in spatial patterns among fish populations in Victoria's Port Phillip Bay. Measurements on fish biomass for a large number of species were recorded from a variety of sites throughout the Bay. A matrix of 'similarities' between pairs of species–site data was obtained and this was analysed by conventional MDS methods. Figure A5.22 reveals three distinct groupings or clusters of observations that the researchers were able to identify with different habitat classifications (pale grey = shallow; grey = medium; dark grey = deep). In this case, the MDS has achieved its objective of extracting a pattern from data that otherwise would be difficult to visualise in many dimensions.

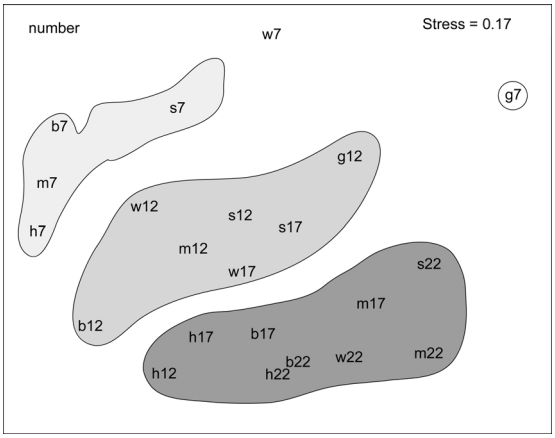


Figure A5.22. MDS plot of fish habitat data in Port Phillip Bay (taken from Parry et al. 1996)

There are differing views about the extent to which MDS can be used as an inferential tool. Computationally intensive methods are available which enable the researcher to conduct formal tests of significance, although some would argue that the strength of MDS lies in its descriptive capability and that it should thus be confined to the realm of exploratory data analysis (EDA).

A5.2. Worked Examples (see next 15 pages)

The worked examples were prepared using MINITAB® statistical software. These examples have been adapted from real analyses. They include actual computer output, represented by typewriter font.

Worked Example 1: Checking Distributional Assumptions

Context

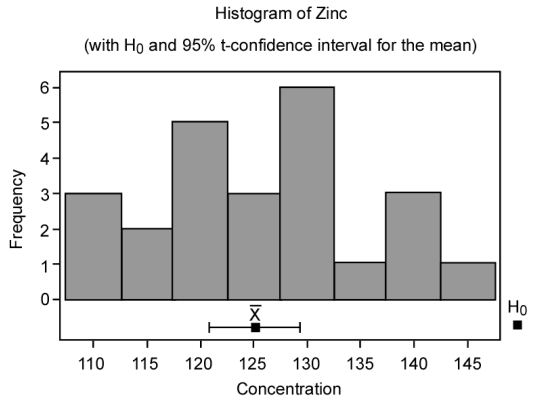
There have been concerns in recent years over the effects that contaminants in our waterways are having on marine life. Organisms such as fish and oysters are known to accumulate heavy metals in their tissues to such dangerously high levels that they can be unsafe to eat. This is especially the case in the Georges River, Sydney. Therefore a research scientist was interested to see if the zinc concentration in fish from the Georges River were at a concentration that would be harmful to the health of humans; 24 fish were caught and the zinc concentrations were determined. The results were as follows:

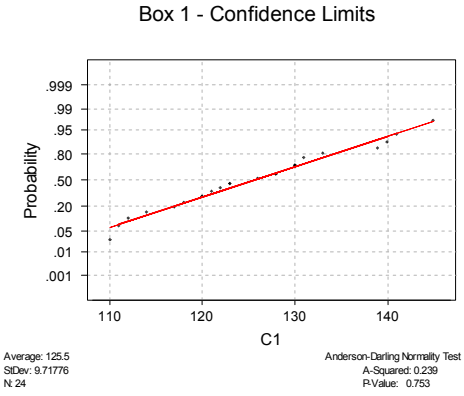
Zn(μg/g)			
141	130	130	112
126	123	133	131
128	139	130	145
128	122	114	117
110	118	111	120
121	140	123	120

Analysis

A useful way of checking on the distributional assumption (normal or otherwise) is via a *probability plot*.

The normal probability plot for the Zn data is shown below. Departures from linearity provide evidence to suggest non-normality. Various formal statistical tests of the hypothesis of normality can also be conducted. Results from the Anderson–Darling test of this hypothesis confirm the impression given by the plot — that the assumption of a normally distributed parent population is tenable in this instance.





The sample histogram is shown above (not overly informative for a small sample size) with an indication of the relative positions of the sample mean, the 95% confidence interval, and the value hypothesised under H_0 (in this case 150 μg/g corresponding to the NFA guideline). In this case, the hypothesised value of 150 is well removed from the extremities of the confidence interval. A formal test of the null hypothesis is also given below:

t-Test of the Mean (see Table 6.9(a) test for a single population mean)

```
Test of mu = 150 vs mu > 150

Variable      95.0% Lower Bound        T      P
Zinc                     122.10   -12.35  1.000
```

Worked Example 2: Two-sample t-test

Context

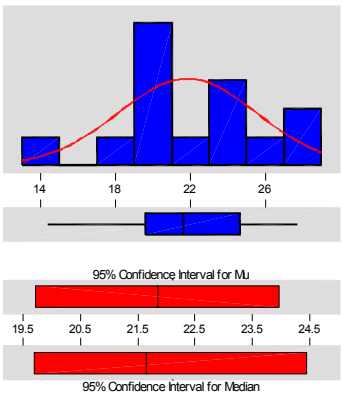
A researcher was interested to know if there was a difference in cadmium concentration between male and female oysters (*Saccostrea crassostrea*). Sample oysters, 13 males and 14 females, were obtained from the Clyde River, Batemans Bay, and analysed for cadmium. The results (in µg/g) were as follows:

Male		Female	
21.4	23.7	27.1	19.7
22.3	21.9	19.4	19.8
28.4	28.2	22.5	27.7
21.7	22.2	25.4	19.9
29.0	17.9	20.8	17.9
24.5	11.7	23.6	14.4
21.4		23.2	24.4

Preliminary Analysis

Descriptive methods (graphical, tabular, and summary) are important tools for teasing out important properties of sample data. The choice of statistical tools has little or nothing to do with the assumed underlying distribution. The panels below show summary presentations of Cd concentrations in male and female oysters.

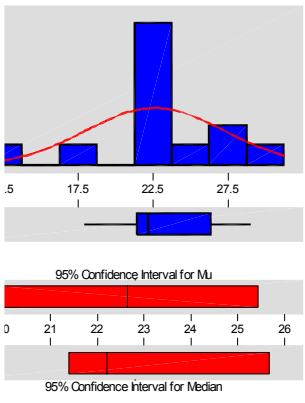
Descriptive Statistics



Variable: Cd - Fem

Anderson-Darling Normality Test	
A-Squared:	0.215
P-Value:	0.811
Mean	21.8429
StDev	3.6903
Variance	13.6180
Skewness	-1.4E-01
Kurtosis	-8.6E-01
N	14
Minimum	14.4000
1st Quartile	19.6250
Median	21.6500
3rd Quartile	24.6500
Maximum	27.7000
95% Confidence Interval for Mu	
19.7122	23.9736
95% Confidence Interval for Sigma	
2.6753	5.9452
95% Confidence Interval for Median	
19.6846	24.4515

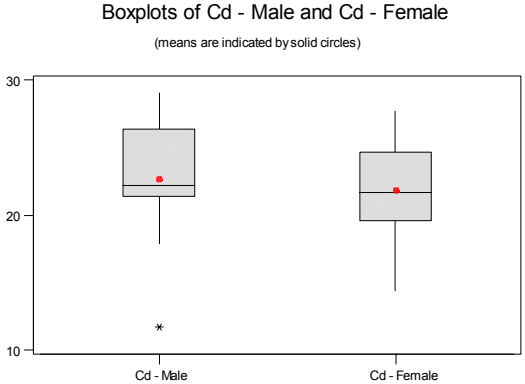
Descriptive Statistics



Variable: Cd - Mal

Anderson-Darling Normality Test	
A-Squared:	0.587
P-Value:	0.101
Mean	22.6385
StDev	4.6196
Variance	21.3409
Skewness	-6.2E-01
Kurtosis	0.130699
N	13
Minimum	11.7000
1st Quartile	21.4000
Median	22.2000
3rd Quartile	26.3500
Maximum	29.0000
95% Confidence Interval for Mu	
19.8468	25.4301
95% Confidence Interval for Sigma	
3.3127	7.6258
95% Confidence Interval for Median	
21.4000	25.6669

Another useful graphical device for summarising and comparing data sets is the box-plot. Sample box-plots for the male and female Cd data are illustrated below. Although the means (and medians) are very similar, the distribution of Cd in the males is (positively) skewed.



Worked Example 2 (continued)

Formal Analysis

Two Sample *t*-Test and Confidence Interval

```
Two sample t for Cd - Male vs Cd – Female
           N      Mean     StDev   SE Mean
Cd - Mal  13     22.64      4.62       1.3
Cd - Fem  14     21.84      3.69      0.99

95% CI for mu Cd - Mal - mu Cd - Fem: ( -2.6,  4.15)
t-Test mu Cd - Mal = mu Cd - Fem (vs not =): t= 0.49  P=0.63  DF=  22
```


```
95% CI for mu Cd - Mal - mu Cd - Fem: ( -2.6,  4.15)
t-Test mu Cd - Mal = mu Cd - Fem (vs not =): t= 0.49  P=0.63  DF=  22
```


Interpretation

The first part of the output above gives the relevant statistics for the difference between males' and females' cadmium concentrations. The means, standard deviations, and standard errors are all in close agreement. The significance of the difference between sample means is tested formally (and equivalently) by reference to the next two lines.

The confidence interval approach shows that a 95% confidence interval for the *difference* (male – female) ranges from –2.6 to 4.15. Since this interval includes zero, we cannot rule out a zero difference. That is, the inequality between male and female cadmium concentrations has *not* been established.

The alternative approach is to conduct the two-sample *t*-test and assess the significance of the result using a *p*-value. We see that the computed *t*-value of 0.49 has an associated *p*-value of 0.63. Since this *p*-value represents the probability due to chance of obtaining a *t*-score of 0.49 (or higher) when the null hypothesis is true, we accept the null hypothesis of equality. To reject this hypothesis at a level of significance α, the computed *p*-value would need to be less than or equal to α.

Worked Example 3: Paired *t*-test for Dependent Samples

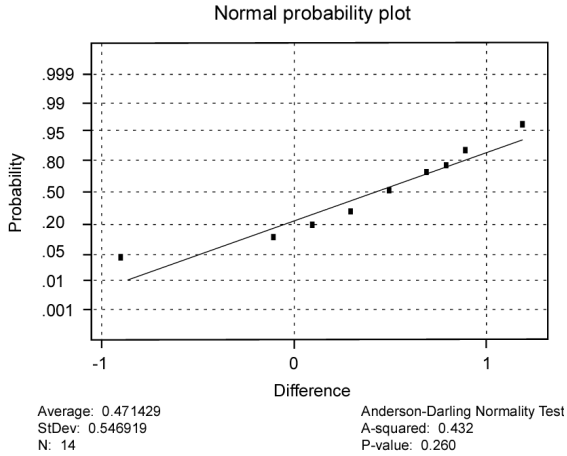
Context

A biologist was interested in determining whether the size of eggs produced by the freshwater turtle *Emydura macquarii* was constant or highly variable between years. Fourteen females from the Macleay River were caught in both 1995 and 1996 and were gravid on each occasion. Females were x-rayed and induced and the eggs were weighed. The average weight of eggs in a clutch for each turtle for the years 1995 and 1996 were as follows:

ID	1995	1996	Difference	ID	1995	1996	Difference
501	7.5	8.3	0.8	536	7.3	6.4	–0.9
504	7.2	7.7	0.5	537	7.6	7.7	0.1
505	6.8	7.7	0.9	543	5.8	6.7	0.9
506	7.3	7.8	0.5	545	7.4	7.9	0.5
524	6.8	7.3	0.5	553	7.4	8.1	0.7
529	8.1	8.0	–0.1	561	7.7	8.0	0.3
530	5.6	6.8	1.2	563	6.5	6.8	0.3

Preliminary Analysis

A preliminary check of the normality assumption for the *differences* (1996 – 1995) is shown below. Neither the plot nor the results of the normality test suggest that the assumption of normally distributed differences has been violated.



Formal Analysis

The test procedure is straightforward and is equivalent to a one-sample *t*-test performed on the differences. The degrees of freedom in this case are the number of *pairs* less one, not the total sample size less one.

t-Test of the Mean

```
Test of mu = 0.000 vs mu not = 0.000

Variable     N      Mean    StDev   SE Mean        t          P
diff        14     0.471    0.547     0.146     3.23     0.0066
```

Interpretation

At a 5% level, the p-value of 0.0066 is deemed to be significant and we therefore reject the hypothesis of equality of the two population means.

Worked Example 4: Fixed Effect ANOVA

Context

A biologist wanted to know if there was a difference in the abundance of the mayfly nymphs (*Ulmerophlebia* sp.) between the northern, southern, eastern, and western shores of Lake Windermere, Jervis Bay. The western shoreline bears the brunt of prevailing winds in the region, and the resulting wave action has reduced emergent macrophytic growth and accumulation of litter on both the western and the northern shores. Five replicate column collections of benthic invertebrates were collected from each of the four shores and the nymphs of the mayfly *Ulmerophlebia* sp. were counted. The data below have been transformed to a log scale to weaken the relationship between the variance and the mean.

West	North	East	South
0.48	1.11	2.00	1.94
0.78	1.15	1.36	1.52
1.11	1.57	1.86	1.60
0.60	0.95	2.33	1.81
1.30	1.48	1.83	2.26

Preliminary Analysis

Descriptive Statistics

```
Variable  direct          N     Mean   Median  Tr Mean    StDev  SE Mean
log(abund)W               5    0.854    0.778    0.854    0.346    0.155
          N               5    1.252    1.146    1.252    0.260    0.116
          E               5    1.876    1.857    1.876    0.349    0.156
          S               5    1.827    1.813    1.827    0.294    0.132

Variable  direct        Min      Max       Q1       Q3
log(abund)W           0.477    1.301    0.540    1.207
          N           0.954    1.568    1.034    1.523
          E           1.362    2.330    1.597    2.165
          S           1.519    2.260    1.560    2.100
```

Observe that the standard deviations are reasonably consistent over the four sites. The main difference in *log*-abundance seems to be depressed counts at the western and northern sites. The significance of these differences is formally examined via a one-way (fixed effects) ANOVA model.

Formal Analysis

```
Analysis of Variance for log(abund)
Source     DF        SS        MS        F        P
direct      3    3.5877    1.1959    12.09    0.000
Error      16    1.5830    0.0989
Total      19    5.1707
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level       N      Mean     StDev  ---------+---------+---------+-------
W           5    0.8545    0.3459  (-----*-----)
N           5    1.2519    0.2596         (-----*-----)
E           5    1.8764    0.3495                      (-----*----)
S           5    1.8266    0.2942                     (-----*----)
                                  ---------+---------+---------+-------
Pooled StDev =   0.3145                 1.00      1.50      2.00
```

Interpretation

The computed *F*-ratio of 12.09 (with 3 and 16 degrees of freedom) is seen to be significant ($p < 0.0005$). Furthermore, the individual confidence intervals (based on the pooled estimate of common variance) suggest that abundances at the eastern and southern sites are similar as are the western and northern sites, but that these two groupings are different. The significance of this observation is examined via Tukey's multiple comparison technique.

Worked Example 4 (continued)

Tukey's pairwise comparisons

```
Family error rate = 0.0500
Individual error rate = 0.0113
Intervals for (column level mean) - (row level mean)
                W           N           E
       N     -0.9672
              0.1722
       E     -1.5916     -1.1942
             -0.4522     -0.0548
       S     -1.5419     -1.1444     -0.5199
             -0.4024     -0.0050      0.6195
```

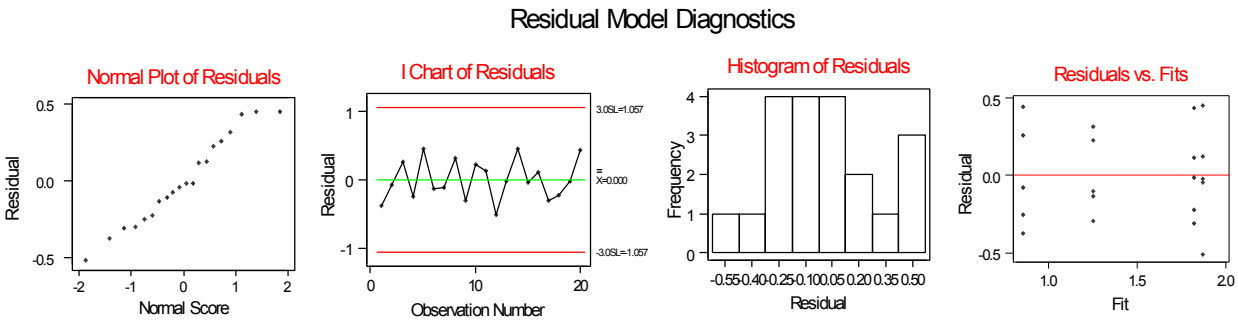
Interpretation

The analysis above shows a matrix of confidence limits for the respective comparison between row and column means. For example, the top left cell shows that the confidence interval for the *difference* between the western and northern site means extends from –0.9672 to 0.1722. Given that equality of the two means corresponds to a *zero* difference we would *not* reject a null hypothesis of no difference between this pair of sites (since zero is spanned by the confidence interval). Similar interpretations may be made for the other cell entries. Overall, our conclusion is as we suspected — there is no difference between North–West or East–South sites, but these groupings are significantly different from each other. The so-called 'Family error rate' is effectively the Type I error for the set of six comparisons — that is, the probability of incorrectly rejecting at least one true null hypothesis of no difference is 0.05. To achieve this overall error rate of 0.05, each individual comparison has a Type I error of 0.0113.

Residual Analysis

It is always instructive to look at the *residuals* (i.e. the portion of the original measurement that the statistical model has not been able to account for) after the ANOVA model has been fitted. A number of diagnostic tools are available to help decide if there have been any violations of the attendant ANOVA assumptions.

The plots below show one such set of diagnostic tools:



The normal probability plot shows some departure from linearity (particularly in the tails of the distribution) suggesting that the normal assumption may have been violated (further testing as in Worked Example 1 would be necessary to establish the significance of these departures). The I-chart shows the residuals plotted in sequential order. Evidence of trends or non-random behaviour is indicative of some underlying process (e.g. seasonal effects) that has not been captured by the model and may invalidate the ANOVA output. Such effects do not appear to be evident from the plot. The histogram provides another visual check on the distribution of residuals. The non-normality is apparent, although the small sample sizes should be kept in mind when assessing the histogram. Given that the normality assumption is a robust assumption, the shape of the histogram of residuals should not overly concern us. Finally, the plot of residuals against the fitted values (i.e. the predicted mean abundance in each of the four directions) should also show no systematic trends, drifts, periodicities or other non-random behaviour. We should also look for evidence of gross discrepancies in the degree of spread between the different factor levels as evidence of non-constant variance. In the example above, no particular concerns are raised.

Worked Example 5: Single Factor ANOVA Planned Comparison

Context

Phosphorus is an important nutrient in aquatic ecosystems; concentrations can be changed dramatically through non-natural discharges into streams and lakes. The following data are for concentrations of phosphorus (μg/L) in samples of water taken at various distances up- and downstream of a waste water outlet. The data below are replicates taken from their respective locations at the one time:

Distance downstream (km)				
–0.5	0.0	1.0	2.0	3.0
4.86	6.16	6.82	5.86	5.31
4.86	5.83	6.67	5.73	4.98
5.19	6.93	6.34	5.62	4.98
4.31	6.16	6.08	4.83	5.46
4.99	6.93	5.73	5.49	4.66

The water scientist wanted to know whether the mean phosphorus concentrations of sites downstream from the effluent outlet differed from the upstream site ('control'). The scientist also needed to assess if the impact, if any, could be considered local or if it persisted well downstream.

Analysis

Analysis of Variance for P

```
Source     DF        SS        MS        F        P
Dist        4    10.121     2.530    15.59    0.000
Error      20     3.246     0.162
Total      24    13.367
```

```
                                Individual 95% CIs For Mean
                                Based on Pooled StDev
Level       N      Mean     StDev  -------+---------+---------+---------
-0.5        5    4.8420    0.3266   (----*-----)
 0.0        5    6.4020    0.5005                        (----*-----)
 1.0        5    6.3280    0.4411                       (----*-----)
 2.0        5    5.5060    0.4018            (-----*----)
 3.0        5    5.0780    0.3137      (-----*----)
                                -------+---------+---------+---------
Pooled StDev =   0.4029              4.90      5.60      6.30
```

Interpretation

The ANOVA output suggests a highly significant 'distance' effect — in other words, phosphorus concentrations are related to distance from outlet. Examination of the individual confidence intervals suggests upstream concentrations are reached beyond 2 km downstream from the outlet. Marked differences in concentrations are observed at the outlet and 1 km downstream.

A formal test of the significance of these observations can be made through the use of an appropriate multiple comparison procedure. Tukey's method was used in Worked Example 4, although the situation here is slightly different because one of the locations is a control with which all other sites are to be compared. The appropriate procedure in this case is Dunnett's test. Like Tukey's test, Dunnett's test keeps the overall experiment-wise error rate fixed at the nominal 0.05 level by conducting each of the individual comparisons at an appropriately smaller level of significance.

Dunnett's intervals for treatment mean minus control mean

```
Family error rate = 0.0500
Individual error rate = 0.0153
Control = level (-0.5) of Dist
Level        Lower    Centre     Upper ------+---------+---------+---------+-
 0.0        0.8848    1.5600    2.2352                  (-------*--------)
 1.0        0.8108    1.4860    2.1612                 (--------*-------)
 2.0       -0.0112    0.6640    1.3392       (-------*--------)
 3.0       -0.4392    0.2360    0.9112  (-------*-------)
                                      ------+---------+---------+---------+-
                                         0.00      0.80      1.60      2.40
```

Our intuition is supported by the results of Dunnett's test. We see that the control site is significantly different from the 0.0 and 1.0 downstream sites (since zero is not encompassed by the relevant interval) but is not significantly different from the 2.0 or 3.0 downstream sites.

Worked Example 6: Nested Analysis of Variance

Context

A water scientist was interested to know if there was a difference in fluoride concentration between three locations. Three independent water samples were taken from each location. Two independent determinations of fluoride content (mg/L) were made on each sample. The data were as follows (data from Zar 1984):

Locations	1			2			3		
Samples	1	2	3	1	2	3	1	2	3
	1.1	1.3	1.2	1.3	1.3	1.4	1.8	2.1	2.2
	1.2	1.1	1.0	1.4	1.5	1.2	2.0	2.0	1.9

This is in the design of a nested ANOVA with samples (random factor) nested within location (fixed factor). The inclusion of a random-effects nested factor (sample) enables us to account for some of the within-group variability, and hence improves the sensitivity of the test with respect to location effects.

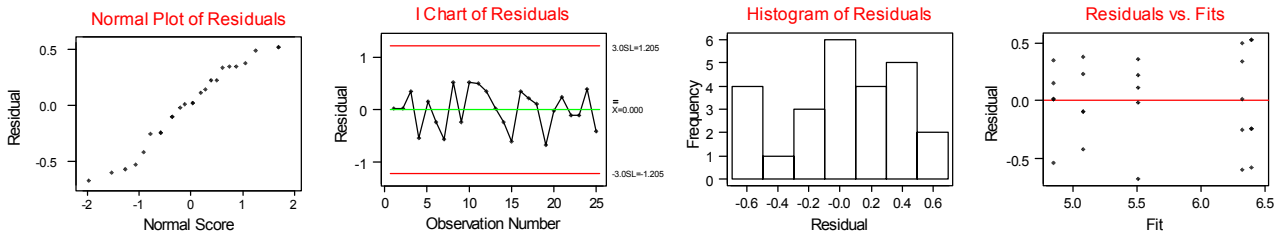
Analysis

This nested ANOVA model is conveniently specified using the GLM (General Linear Model) option available in most statistical software packages. The following output is obtained.

General Linear Model

```
Factor              Type  Levels   Values
Location           fixed       3   1 2 3
Sample(Location) random       9   1 2 3 1 2 3 1 2 3

Analysis of Variance for Concen, using Adjusted SS for Tests

Source              DF     Seq SS     Adj SS     Adj MS       F      P
Location             2    2.37000    2.37000    1.18500  142.20  0.000
Sample(Location)     6    0.05000    0.05000    0.00833    0.47  0.816
Error                9    0.16000    0.16000    0.01778
Total               17    2.58000

Variance Components, using Adjusted SS

Source            Estimated Value
Sample(Location)          -0.00472
Error                      0.01778
```

An analysis of the residuals after fitting the nested ANOVA model is provided below.



Interpretation

The ANOVA output suggests a highly significant 'location' effect, although samples within location do not differ significantly.

With respect to the analysis of residual diagnostics: there is some departure from normality (although not serious) in the lower tail of the distribution (refer normal probability plot and histogram). The plot of residuals versus fitted values shows some evidence of non-constant variance. Formal tests of the homogeneity of variance assumption are available (see Worked Example 9, for instance).

Worked Example 7: Equality of Variances

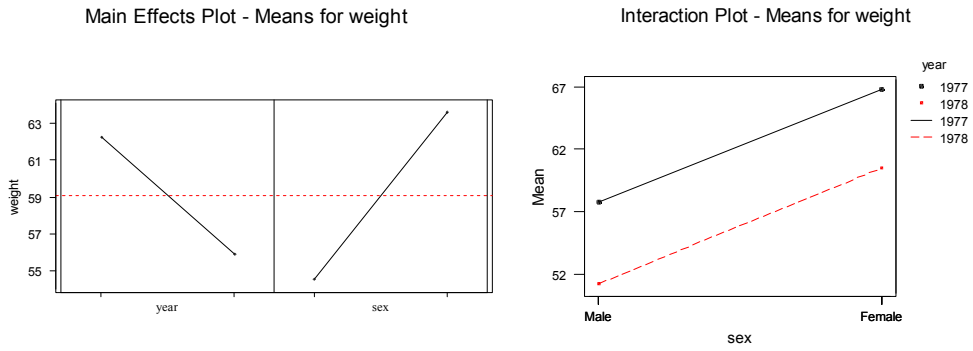
Context

A researcher was interested in determining if there was a difference in body weight of *Antechinus stuartii* between the years 1977 and 1978. The researcher knew that females were bigger than males so the sex of each individual was also recorded. The data were as follows:

	1977	1978		1977	1978
	56	52		61	54
Male	56	52	Female	63	64
	62	52		72	64
	57	49		71	60

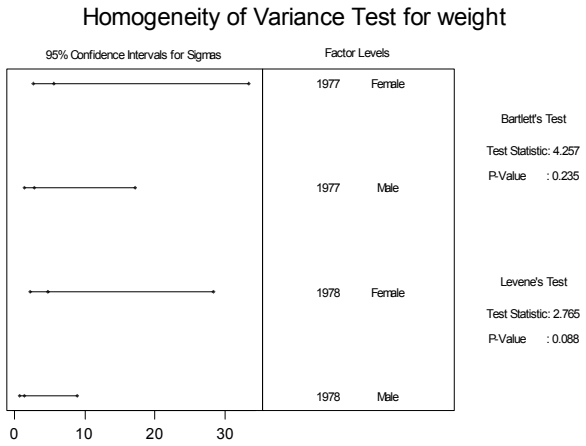
Analysis

It is always useful to inspect main effects and interaction plots in multi-factor designs to gain a better appreciation of the results of any subsequent ANOVA / GLM analysis.



The plots above suggest the presence of both a 'year' and a 'sex' main effect. The interaction plot shows two almost parallel lines that are indicative of an absence of any interaction between the main effects. The significance of these observations can be formally tested via a two-factor ANOVA.

Perhaps more important than the assumption of normality, is the ANOVA assumption of equality of group variances. It is good practice to examine the group statistics and make some preliminary assessment of the applicability of the homogeneity of variances assumption. More formal statistical tests are available, two of which (Bartlett's and Levene's test) are reproduced below. It should be noted that one drawback of Bartlett's test is its sensitivity to departures from normality. A more robust test is Hartley's F_{max} test (not conducted here).



Interpretation

In this case, we accept the hypothesis of equal group variances at the 5% level, although the wide discrepancy in *p*-values between the two test procedures is noted. Good statistical practice would dictate that the most appropriate test is identified and then applied to the data. To apply a battery of tests and choose the most appealing result is of course unscientific and negates the application of any statistical method.

Worked Example 8: Two-factor ANOVA

Context

As for Worked Example 7.

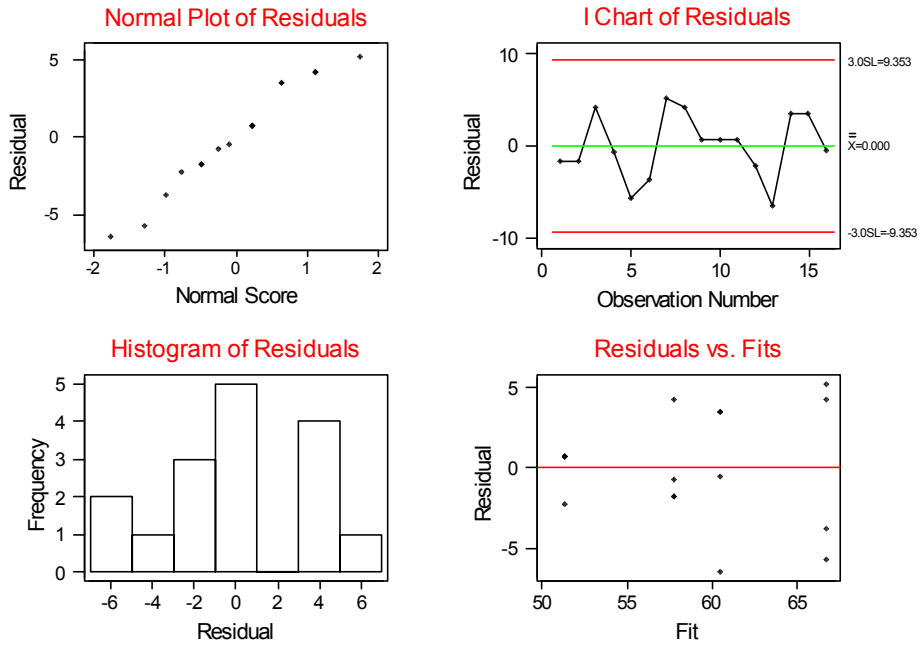
Analysis

A fixed two-factor analysis of variance is appropriate as it enables the comparison of weight between years after adjusting for differences between sexes. Conversely sex can also be compared after adjusting for differences between years.

```
Analysis of Variance for weight

Source      DF     Seq SS     Adj SS     Adj MS       F      P
year         1     162.56     162.56     162.56   10.20  0.008
sex          1     333.06     333.06     333.06   20.90  0.000
year*sex     1       0.06       0.06       0.06    0.00  0.951
Error       12     191.25     191.25      15.94
Total       15     686.94
```

Interpretation

This analysis confirms our observations. Both main effects are significant at a 5% level, while the interaction is non-significant. The residual diagnostics appear below.



The residual diagnostics give no cause for concern or remedial action.

Worked Example 9: Analysis of Covariance

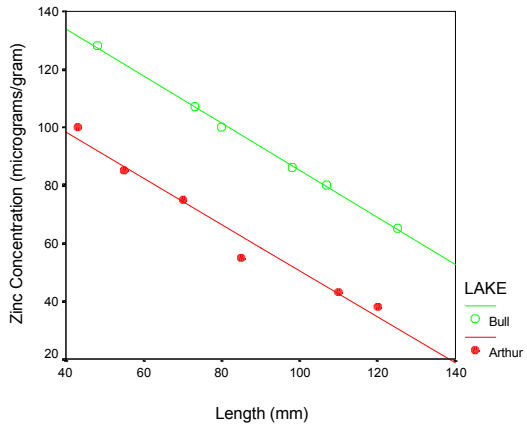
Context

Fisheries were interested to know whether fish in a polluted environment were likely to accumulate higher concentrations of heavy metals than fish in a more pristine habitat. To investigate this, six flathead were caught in both Lake Bull (polluted urban lake) and Lake Arthur (relatively pristine) and analysed for zinc (µg/g). As heavy metal concentration in organisms is often related to body size, the lengths of the fish were also recorded. The data were as follows:

Site	Zinc	Length	Site	Zinc	Length
Lake Arthur	43	110	Lake Bull	80	107
Lake Arthur	75	70	Lake Bull	86	98
Lake Arthur	38	120	Lake Bull	65	125
Lake Arthur	55	85	Lake Bull	100	80
Lake Arthur	100	43	Lake Bull	128	48
Lake Arthur	85	55	Lake Bull	107	73

As there was large variability in the size of fish sampled, it was decided that comparisons of zinc concentrations between lakes should be made by analysis of covariance (ANCOVA). The ANCOVA tests for differences in zinc concentration in fish between lakes after correcting the zinc measurements for the effect of fish body size.

Analysis



- We first determine that the relationship between the dependent variable (Zinc) and the covariate (fish length) is linear for both lakes.
- From the diagram it is clear that there is a strong negative linear relationship between zinc concentration and the length of fish for both Lake Bull and Lake Arthur. Therefore, zinc concentration is dependent on fish size in both lakes.

The zinc data are analysed using an ANCOVA model with length being a covariate. We note that two model formulations are possible, depending on the inference to be made about the covariate.

In the first formulation we use a nested model to force the explicit estimation of separate regression slopes (i.e. one regression for each of the two lakes).

Nested Analysis of Variance for Zinc

```
Source          DF     Seq SS     Adj SS     Adj MS       F      P
Lake             1     2408.3      337.3      337.3   30.15  0.000
Length(Lake)     2     5403.9     5403.9     2701.9  241.57  0.000
Error            8       89.5       89.5       11.2
Total           11     7901.7

Term                 Coef     StDev        T      P
Constant          148.298     3.278    45.24  0.000
Length (Lake)
       Arthur   -0.79870   0.04922   -16.23  0.000
       Bull     -0.81318   0.05485   -14.83  0.000
```

Worked Example 9 (continued)

```
Unusual Observations for Zinc

Obs      Zinc       Fit   StDev Fit  Residual   St Resid
  4    55.000    62.406       1.383    -7.406     -2.43R

R denotes an observation with a large standardised residual
```

From this analysis we conclude that (i) zinc concentrations are significantly different between the two lakes, and (ii) there is a significant relationship (the implied null hypothesis is that of zero slope for the regression) between zinc levels and length in each of the two lakes. The computer output has also flagged a potential outlier or in some other way aberrant observation (Observation #4). The high standardised residual for this observation is a consequence of the extremely high zinc reading.

The next model formulation is a fully crossed design in which the overall regression effect is estimated and its significance tested.

Fully-crossed Analysis of Variance for Zinc

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Lake	1	2408.3	337.3	337.3	30.15	0.000
Length	1	5403.4	5350.8	5350.8	478.39	0.000
Lake*Length	1	0.4	0.4	0.4	0.04	0.849
Error	8	89.5	89.5	11.2		
Total	11	7901.7				

Term		Coef	StDev	T	P
Constant		148.298	3.278	45.24	0.000
Length		-0.80594	0.03685	-21.87	0.000
Length*	Lake				
	Arthur	0.00724	0.03685	0.20	0.849

Interpretation

From this analysis we conclude that the overall coefficients of the regression slope and intercept are both highly significant (i.e. non-zero). Furthermore, the interaction term is non-significant suggesting that separate slopes are not warranted (implying essentially parallel lines).

Worked Example 10: Generalised Linear Models

Context

We note that the data analysed in this example take the form of counts, which are necessarily discrete. By back-transforming, the original set of abundances is obtained. These are shown below:

Variable	direct	N	
Count	W	5	3, 6, 13, 4, 20
	N	5	13, 14, 37, 9, 30
	E	5	100, 23, 72, 214, 68
	S	5	87, 133, 40, 65, 182

It is evident that there are not only substantial differences *between* directions, but also of the abundances *within* each direction. Generally, normal-based models are satisfactory for the analysis of count data provided the average counts are reasonably large (typically > 30). While this is mostly true for the data above, some quite low counts are observed in the north and westerly directions.

Analysis

A more flexible, and conceptually more appropriate approach to the analysis of this type of (discrete) data is afforded by a *generalised linear model* having a Poisson error term and *log-link*. The details of this approach are not covered here, although the analysis is similar to a conventional ANOVA with the exception that we look at changes to the *deviance* statistic as a series of nested models is fitted.

Interpretation

For the original data, we find the null model (corresponding to a single overall mean) has a deviance of 1019.3 with 19 degrees of freedom. By adding the 'direction' factor to the model, this deviance is reduced to 411.71 (16 degrees of freedom). The change in deviance 607.59 is highly significant when compared to a chi-squared test statistic having 3 (19–16) degrees of freedom.

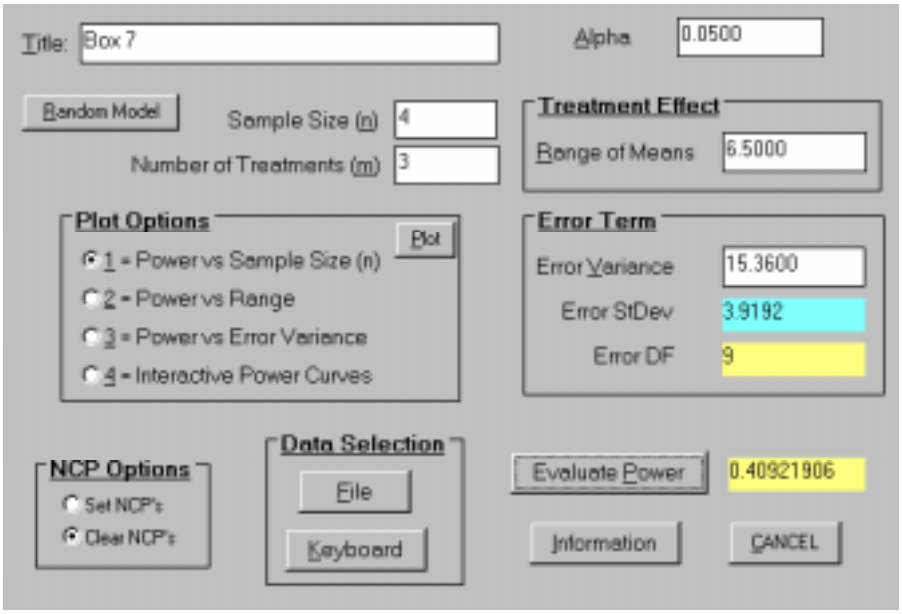
Worked Example 11: Power and Sample Size Determinations

Context

A pilot study was carried out to determine the mean cadmium concentrations in the cockle *Anadara trapezia* from each of three sites. Estimates of means for each site from four samples were: A: 12.75; B: 17.75; C: 19.25, and σ^2 was estimated to be 15.36. What will be the power of the ANOVA in detecting a difference between sites if we test at the 0.05 level of significance?

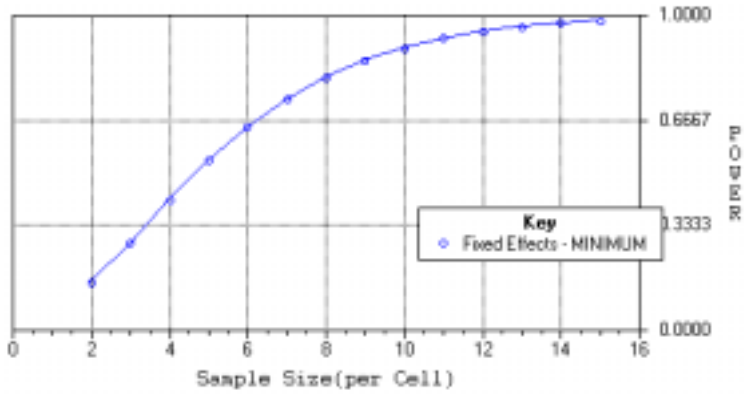
Analysis

Output from the *PowerPlant*® software (see section A5.1.10 for download details) is shown below.



Interpretation

We see immediately from the 'Evaluate Power' box, that the estimated power for this design is 0.409. Sample size determinations are readily assessed from the power curve.



We see that a sample size of about 13 per treatment group should give the required 0.8 power.

Worked Example 12: Regression

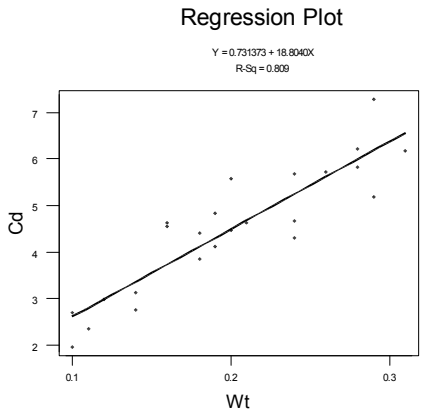
Context

A marine scientist wanted to determine if the cockle *Anadara trapezia* was an appropriate bioindicator for heavy metal pollution. Other species of bivalves have previously been found to accumulate heavy metals in their body tissues. Therefore, 24 cockles were dried, weighed, and the cadmium determined.

Cd	Wt	Cd	Wt	Cd	Wt	Cd	Wt
2.71	0.10	2.35	0.11	1.97	0.10	4.67	0.24
2.76	0.14	4.46	0.20	4.54	0.16	3.14	0.14
4.12	0.19	4.64	0.16	4.41	0.18	5.57	0.20
4.83	0.19	3.85	0.18	5.68	0.24	5.72	0.26
6.16	0.31	6.21	0.28	4.30	0.24	4.64	0.21
7.27	0.29	3.00	0.12	5.82	0.28	5.18	0.29

Issue to be addressed: Is there a relationship between the size of the cockle and the total amount of cadmium in the tissues? A positive relationship would suggest that cockles do in fact accumulate cadmium in their tissues over time.

Analysis

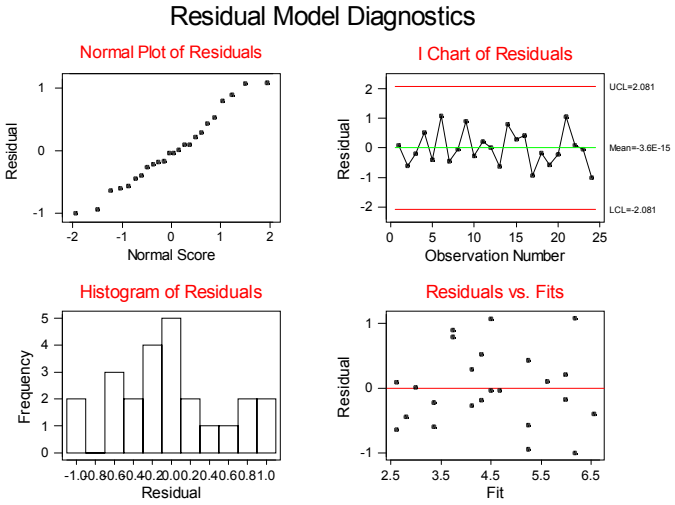



```
The regression equation is
y = 0.731 + 18.8 x

Predictor       Coef     StDev         T       P
Constant      0.7314    0.4091      1.79   0.088
x             18.804     1.947      9.66   0.000

S = 0.6018     R-Sq = 80.9%     R-Sq(adj) = 80.0%
```

Interpretation

Both the plot and the analysis indicate a strong, positive linear relationship between Cd and weight. Inspection of the *p*-values associated with terms in the regression model suggests that the intercept (0.7314) is not significantly different from zero (hence a regression model through the origin might be contemplated) while the regression slope (18.804) is significantly different from zero.

An important follow-up analysis relates to the regression diagnostics. These appear below. These diagnostic plots raise no particular concerns.



Appendix 6

Typical Field Record Sheet & Laboratory Request Form

A6.1. Field Record Sheet

Officer/s……..…….....……. **Date**……

Sampling run number:……..... **Site code**:…….....…….

Site name…….....…

Time: start..... **finish**.....

Field measurements:

Parameter	Result
Depth (m)	
Secchi depth (m)	
Altitude (m)	
Temperature (^{o}C)	
Turbidity (NTU)	
Dissolved oxygen (mg/L) (% saturation)	
Electrical conductivity (mS/cm)	
pH	
Salinity ( )	
Eh (mV)	
Others	

Field observations:

Station no.…….

Description.....

.....

Observation	Details
Weather: e.g. wind, wind direction, cloud cover	
Colour and appearance of water	
Water surface condition	
Water flow, level, tide:	
Presence of nuisance organisms (e.g. macrophytes, phytoplankton scums, algal mats)?	
Presence of oily films on surface or on shoreline?	
Presence of floating debris or grease?	
Presence of odour or frothing?	
Other observations	

Signature.....

(when sample collected and entries completed)

Water quantity measurement data:

Location description.....

Description of gauge.....

Stage height.....

Time.....

Sample details:

Analyte	Container material	Volume collected	Preservation	Quality control
Major ions				
Metals				
Organic compounds				
Pesticides and herbicides				
Mercury				
Phenols				
Nutrients				
BOD and COD				
Others				

Quality Control Remarks:.....

.....

.....

.....

.....

.....

.....

A6.2. Typical Laboratory Request Form

(partially filled-in as an example)

Sample program / site description.....

Sampling officer.....**Title**.....**Section**.....**Branch**.....

Sampling date.....**File Reference**.....**Payment Authority No.**.....

Sample no.	Sample time	Sample location	Sample description	Parameters to be analysed	Sample size (mL or g)	Reference criteria
MB1	09.00	Monitor bore (north-east)	Groundwater	pH, EC, TP, TN	1000 mL	SW, IP
SP2	09.45	Creek upstream of site	Creek water	Colour, NFR, EC, TP, TN	1000 mL	A, AE

Legend for reference criteria: AE = aquatic ecosystems, **DW** = drinking water, **IP** = irrigation of plants, **SW =** stock water supplies, **IW** = industrial water supplies, **A** = aesthetic values; **R =** Recreational waters (reference: ANZECC *Australian & NZ Water Quality Guidelines for Fresh and Marine Waters* (1992))

Site conditions during sampling.....
(e.g. Cool 12^{o}C, raining, creek flowing at estimated 10–15 L/sec)

Sample preservation details.....(e.g. On ice)

Analytical laboratory.....**Accepted by**.....**Title**.....

Date.....**Time**.....

Comments:

Analysis request notes (tick requirement)**:**	**Return analysis results to:**.....
Routine..... (.....) — mail results when ready Urgent..... (.....) — fax results to ....., as soon as possible Legal action.....(.....) — ensure chain of custody and data validation	Entity..... Address..... ....