Intergovernmental
Oceanographic
Commission

Manuals and Guides **34**

ENVIRONMENTAL DESIGN AND ANALYSIS IN MARINE ENVIRONMENTAL SAMPLING

1996 UNESCO

IOC Manuals and Guides

No.	Title
1 rev. 2	Guide to IGOSS Data Archives and Exchange (BATHY and TESAC). 1993. 27 pp. (English, French, Spanish, Russian)
2	International Catalogue of Ocean Data Station. 1976. *(Out of stock)*
3 rev. 2	Guide to Operational Procedures for the Collection and Exchange of IGOSS Data. Second Revised Edition, 1988. 68 pp. (English, French, Spanish, Russian)
4	Guide to Oceanographic and Marine Meteorological Instruments and Observing Practices. 1975. 54 pp. (English)
5	Guide to Establishing a National Oceanographic Data Centre. 1975. *(Out of stock)*
6 rev.	Wave Reporting Procedures for Tide Observers in the Tsunami Warning System. 1968. 30 pp. (English)
7	Guide to Operational Procedures for the IGOSS Pilot Project on Marine Pollution (Petroleum) Monitoring. 1976. 50 pp. (French, Spanish)
8	*(Superseded by IOC Manuals and Guides No. 16)*
9 rev.	Manual on International Oceanographic Data Exchange. (Fifth Edition). 1991. 82 pp. (French, Spanish, Russian)
9 Annex I	*(Superseded by IOC Manuals and Guides No. 17)*
9 Annex II	Guide for Responsible National Oceanographic Data Centres. 1982. 29 pp. (English, French, Spanish, Russian)
10	*(Superseded by IOC Manuals and Guides No. 16)*
11	The Determination of Petroleum Hydrocarbons in Sediments. 1982. 38 pp. (French, Spanish, Russian)
12	Chemical Methods for Use in Marine Environment Monitoring. 1983. 53 pp. (English)
13	Manual for Monitoring Oil and Dissolved/Dispersed Petroleum Hydrocarbons in Marine Waters and on Beaches. 1984. 35 pp. (English, French, Spanish, Russian)
14	Manual on Sea-Level Measurements and Interpretation. 1985. 83 pp. (English, French, Spanish, Russian)
15	Operational Procedures for Sampling the Sea-Surface Microlayer. 1985. 15 pp. (English)
16	Marine Environmental Data Information Referral Catalogue. Third Edition. 1993. 157 pp. (Composite English/French/Spanish/Russian)
17	GF3: A General Formatting System for Geo-referenced Data
	Vol. 1: Introductory Guide to the GF3 Formatting System. 1993. 35 pp. (English, French, Spanish, Russian)
	Vol. 2: Technical Description of the GF3 Format and Code Tables. 1987. 111 pp. (English, French, Spanish, Russian)
	Vol. 4: User Guide to the GF3-Proc Software. 1989. 23 pp. (English, French, Spanish, Russian)
	Vol. 5: Reference Manual for the GF3-Proc Software. 1992. 67 pp. (English, French, Spanish, Russian)
	Vol. 6: Quick Reference Sheets for GF3 and GF3-Proc. 1989. 22 pp. (English, French, Spanish, Russian)
18	User Guide for the Exchange of Measured Wave Data. 1987. 81 pp. (English, French, Spanish, Russian)
19	Guide to IGOSS Specialized Oceanographic Centres (SOCs). 1988. 17 pp. (English, French, Spanish, Russian)
20	Guide to Drifting Data Buoys. 1988. 71 pp. (English, French, Spanish, Russian)
21	*(Superseded by IOC Manuals and Guides No. 25)*
22	GTSPP Real-time Quality Control Manual. 1990. 122 pp. (English)
23	Marine Information Centre Development: An Introductory Manual. 1991. 32 pp. (English, French, Spanish, Russian)
24	Guide to Satellite Remote Sensing of the Marine Environment. 1992. 178 pp. (English)
25	Standard and Reference Materials for Marine Science. Revised Edition. 1993. 577 pp. (English)
26	Manual of Quality Control Procedures for Validation of Oceanographic Data. 1993. 436 pp. (English)
27	Chlorinated Biphenyls in Open Ocean Waters: Sampling, Extraction, Clean-up and Instrumental Determination. 1993. 36 pp. (English)
28	Nutrient Analysis in Tropical Marine Waters. 1993. 24 pp. (English)
29	Protocols for the Joint Global Ocean Flux Study (JGOFS) Core Measurements. 1994. 178 pp . (English)
30	MIM Publication Series:
	Vol. 1: Report on Diagnostic Procedures and a Definition of Minimum Requirements for Providing Information Services on a National and/or Regional Level. 1994. 6 pp. (English)
	Vol. 2: Information Networking: The Development of National or Regional Scientific Information Exchange. 1994. 22 pp. (English)
	Vol. 3: Standard Directory Record Structure for Organizations, Individuals and their Research Interests. 1994. 33 pp. (English)
31	HAB Publication Series:
	Vol. 1: Amnesic Shellfish Poisoning. 1995. 18 pp. (English)
32	Oceanographic Survey Techniques and Living Resources Assessment Methods. 1996. 34 pp. (English)
33	Manual on Harmful Marine Microalgae. 1995. (English)
34	Environmental Design and Analysis in Marine Environmental Sampling. 1996. 86 pp. (English)

Intergovernmental
Oceanographic
Commission

Manuals and Guides

34

ENVIRONMENTAL DESIGN AND ANALYSIS IN MARINE ENVIRONMENTAL SAMPLING

by Professor A.J. Underwood

Institute of Marine Ecology
Marine Ecology Laboratories All
University of Sydney
NSW 2006, Australia

1996 UNESCO

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariats of UNESCO and IOC concerning the legal status of any country, territory, or its authorities, or concerning the delimitations of the frontiers of any country or territory.

For bibliographic purposes this document should be cited as follows:

Underwood A.J.
Environmental Design and Analysis
in Marine Environmental Sampling
IOC *Manuals and Guides No.* 34. UNESCO 1996
(English)

Published in 1996
by the United Nations Educational,
Scientific and Cultural Organization,
7, place de Fontenoy, 75700 Paris

Printed in UNESCO's Workshops

Preface

The IOC-IMO-UNEP/GIPME Groups of Experts on Effects of Pollution (GEEP) has been working for a number of years on promoting new ways of understanding how pollutants affect marine biological systems. A major initial focus of GEEP was on calibration workshops where different methods were tested against one another. The Oslo (1986), Bermuda (1988) and Bremerhaven (1992) workshop publications are widely regarded as benchmarks demonstrating that biological effects methods are reliable tools for measurement of the effects of pollutants discharged to the marine environment. IOC through GEEP, in cooperation with UNEP, has published a series of manuals based on the successful techniques and these are listed at the back of this volume.

Monitoring programmes for chemical contamination and for biological effects of these contaminants are used globally. Yet often the sampling design of such programmes has received little attention. Monitoring programmes are often conducted in order to be able to tell politicians and managers whether or not the quality of a given sea area is improving or getting worse. More often than not the answer, frequently after many years of measurement, is that the trends are difficult to detect. It is no exaggeration to say that countless millions of dollars are wasted in poor sampling design where there is no possibility of getting the answers to the questions posed by the managers and politicians. Sampling design is a key but neglected aspect of chemical and biological effects monitoring.

In this manual, GEEP Vice Chairman, Professor A.J. Underwood of the University of Sydney gives a clear and important account of the key elements of good sampling design, It is our hope that this manual will help change the way that managers and scientists consider their monitoring programmes and that there will be a radical change in sampling design as a consequence.

John S. Gray
University of Oslo
17.03.96

TABLE OF CONTENTS

1. INTRODUCTION

Sampling and monitoring to detect impacts of human activities on marine habitats require very careful thought. Much money, time and very substantial resources are often wasted because data are collected without carefully designing sampling so that statistical analyses and interpretation are valid and reliable.

Several related issues need to be considered, such as the purpose of sampling, the scale and time-course of the problem, the available resources, how long must sampling continue. These and related topics must all be carefully assessed.

Nearly every environmental problem is unique and therefore no step-by-step account of what to do is possible, They all share three properties: they are complicated, the things to be measured are very variable and there is insufficient time and money to allow collection of enough information.

In these guide-lines, only the major issues of designing the sampling programme will be considered. There are many advanced topics, but the basic things must be considered properly first. Most of the material described here relates to Univariate measurements. These are measures on a single variable (number of fish, concentration of nitrate, proportion of diseased animals) which is being analysed to determine whether there has been an environmental impact or, if there has been one, how large are its effects.

Multivariate methods are not considered. These are methods for analysing numerous variables simultaneously. Their most common use in environmental sampling has been to analyse the numbers of animals and plants found together in an assemblage at disturbed and undisturbed locations, A recent overview of relevant procedures can be found in Clarke (1993).

All procedures of analysis must be based on clear understanding of the hypotheses to be tested, on the variability (or imprecision) of the measurements and on the assumptions necessary for statistical procedures. These are all introduced here. Wherever possible, references are made to relevant literature.

It is often difficult to get training and expertise in experimental design and it is usually very time-consuming to gain experience in the use of sampling and statistical procedures. Even so, there is one very good thing about this subject. If you use commonsense, think carefully and plan everything before you take samples, you will usually find that you can make sense of the information. This introduction to the topic is (o help understand some of the issues about which you need to think.

2. INTRODUCTION TO SAMPLING AND ESTIMATION OF ENVIRONMENTAL VARIABLES

2.1 POPULATIONS, FREQUENCY DISTRIBUTIONS AND SAMPLES

Biological observations (in whatever form the data are collected) are not constant, fixed truths, but vary from place to place and time to time. So, of course, do measurements in other sciences, such as environmental chemistry or physics. Even where intrinsic variability is small or absent (i.e. the measurement is of some physical or chemical constant), the machinery used to make measurements and the observers themselves are not constant. Therefore measurements are always variable.

2.1.1 Variability in Measurements

There are intrinsic and extrinsic reasons why biological observations are variable. Intrinsic reasons include fundamental properties of biological systems. For example, the size of an animal or the rate of growth of a plant are subject to genetic variability from one individual to another. Thus, unless the individuals are identical in all genetically determined processes, there is no possibility that they will be identical,

Second, in addition to innate properties of systems, other processes cause variability, Even genetically identical individuals will not grow at the same rate (and therefore finish up the same size) because they do not encounter nor process identical amounts and quality of food throughout their lives.

The numbers of organisms in different parts of some habitat are not identical because of the processes that disperse them from their parental sources. They cannot possibly arrive in their final destinations in uniform numbers.

Over and above the variability caused by the intrinsic properties and processes affecting natural systems, there are extrinsic causes of variation. Methods of measurement introduce new sorts of variability. For example, the amount of chlorophyll per sample is not measurable without using machinery which requires processes of extraction and preparation that are not constant in their effects on different samples.

The combined effects of intrinsic and extrinsic causes of variation mean that no measurement we can take, or event we can observe, will be a constant, fixed representation of the true value. Measuring something on several individuals will result in different values of the measurement. Measuring something on the same individual several times will often result in different values of the measurement because of variations in the methods of measuring and changes induced or naturally occurring in the individuals while the measurements are made.

2.1.2 Observations and Measurements as Frequency Distributions

The simplest way to show the variability of observations is to use a graphical plot of the frequency distribution of the variable being measured. As an example; consider a very simple distribution of the number of sharks per reef in a large reserve. There are 730 reefs in the reserve of which, at the time examined, 90 have no sharks, 164 have one shark each, 209 have two sharks each and so on up to 12 reefs which each have five sharks. No reefs have more than five sharks. The frequency distribution is shown in Figure la, which depicts the frequency (as number) of reefs with each number of sharks (the X-axis). It is a discrete distribution, meaning that there are no fractional numbers.

A second example is a continuous distribution (Figure lb) - the lengths of fish of a species in an aquiculture pond. There are 3, 000 fish and all of them are measured. Obviously, lengths can vary by infinitesimally small increments, limited only by the smallest sub-divisions on the ruler used to measure them. There is now essentially an infinite number of sizes along the X-axis, so the data are grouped into intervals of sizes, arbitrarily chosen for convenience of plotting the graph (Figure lb).

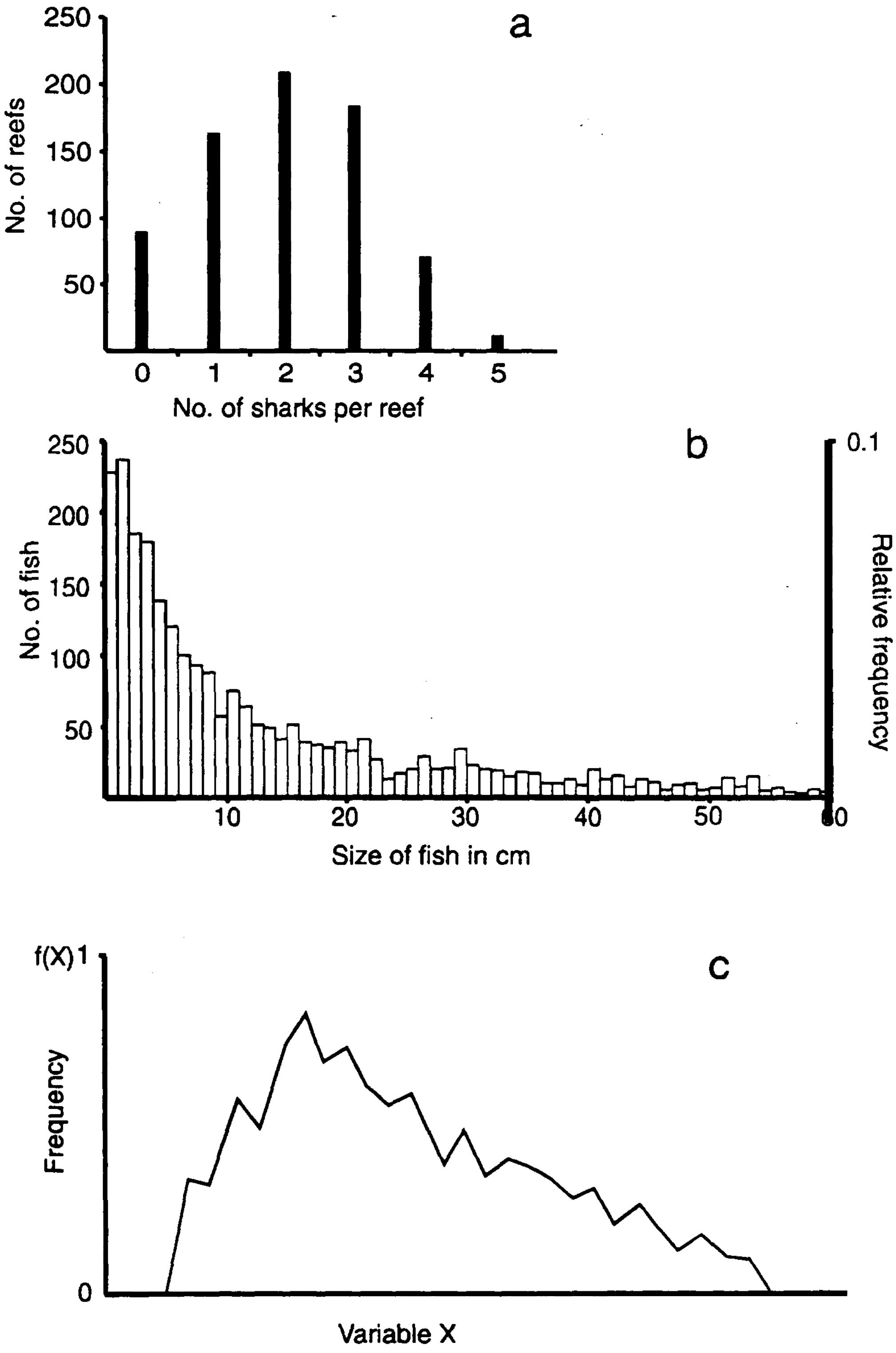

Figure 1. Frequency distributions. (a) discrete distribution of number of 730 reefs with different numbers of sharks. (b) continuous distribution is the sizes of fish in a population of 3000 individuals. (c) the general case of a variable *X* in a population of values, plotted as the proportion of the population having each possible value of *X*.

The general representation of a frequency distribution is shown in Figure 1c, The range of values of the variable (X) is plotted with the frequency ($f(X)$) with which each value of X occurs. Usually, frequency distributions are considered with the frequency plotted as the relative frequency, i.e. the proportion of the entire distribution having each value of the variable (X), rather than as the number of members of the population (e.g. Figure lb). This is convenient because it makes mathematical manipulations of the data simpler and also makes relative frequency distributions from several populations directly comparable, regardless of the sizes of the populations.

2.1.3 Defining the Population to be Observed

The population to be measured or observed must be defined very carefully. This may be simple. For example, measurements of the chlorophyll content of phytoplankton may be needed for a population of phytoplankton in experimental aquaria. On the other hand, it may be more widespread, such as the population of phytoplankton in an ocean. Or the definition may need more information because what is wanted is the chlorophyll content of plankton within 2 km of the coast. The precision of the definition is important for any use of the information gathered.

Suppose that, for whatever reason, a measurement is needed of the size of fish in a population of fish in a particular bay. The fish can only be caught in traps or nets, such that the smallest and fastest-swimming ones will not be caught. The measurements are therefore no longer going to be from the defined population ie all the fish in the bay.

The difficulty comes in when the problem to be solved requires measurements on one population (defined in a particular way) and the measurements are actually only possible in (or are thoughtlessly taken from) a different population (e.g. a subset of the first population).

The population being observed must therefore be defined very precisely to avoid confusion and misunderstanding. If the measurements are taken from a different population from that specified in the problem being examined, they will not be useful for the problem being addressed. It is, at the least, misleading to describe the data as though they came from a population different from the one actually measured.

If measurements must be taken on a population that is not the one specified by the model, hypothesis and null hypothesis being tested, the definition of what was measured becomes very important. A complete description of what was measured is mandatory.

Having defined the relevant and appropriate population to be observed, we now hit the next snag. It is extremely unlikely that the entire population can actually be observed. In the case of determining the chlorophyll content of plankton in an ocean (or part of an ocean), there are so many that it is extremely improbable that you would have the money or the time to measure all of them. In the case of the sharks on a reef, you may be able to search each reef, but the sharks may move around. So, unless you search each reef simultaneously, you cannot know which have and which have not been counted.

Thus, the problem becomes that of being able to measure some variable in a population without knowing the values of the entire population. The problem, as will be developed below, is to determine what sort of subset or sample of the population, out of those elements actually available to you, will best measure the defined variable on behalf of the entire population. Measurements from samples are known as statistics or statistical estimates. Thus, measurements require statistics.

2.1.4 The Location Parameter

The previous sections have introduced the nature of a measurement of a variable. It is clear that to measure something (a variable), you need an estimate of the magnitude of that variable for all elements in the population of interest. The location parameter is the most useful such measure. At this point, it is worth considering the terms variable and parameter in more detail. A variable, as used here so far, is some measurable quantity that differs in magnitude from one member to another in a population, or defined set of such measurements. All the values of the variable in the entire population can be described graphically as a frequency distribution. The shape and range and all other properties of this frequency distribution are determined by its moments. These are mathematical constants that define a frequency distribution. For many distributions (but not all), the moments - the constants defining a distribution - are equal to the parameters (or are simple functions of the parameters) of the distribution. Parameters are constants that define moments; moments define the frequency distribution. In this sense, we are concerned with parameters and their estimation, because these often have empirical interpretative use in biology.

Confusing the two terms variable and parameter has not aided environmental understanding. Phrases such as “the parameter salinity was measured in the river.... ” serve only to confuse things. The variable “salinity” can be measured, i.e. its magnitude observed. The parameters of a distribution of measures of salinity can only be measured if the entire population of all relevant measurements is made. This seems inherently unlikely and, almost certainly, impossible. Common misusage of “parameter” is not improving comprehension. We need the two separate terms “variable” and “parameter”.

The location parameter (L) can be defined operationally as the value which is, on average, as close as possible to all values of variables in the population, The values of a variable in a population can be called X_i(the value for the ith member of the population). There are values $X_1, X_2......X_i......X_N$ where N is the number of members of the population. Therefore, by definition, L must be in the range of all values in the population and must represent the magnitudes of all members of the population. It is the number which is “closest” to all members of the population (i.e. most similar to all Xi values).

L is usually described as the arithmetic average or mean value of the variable in the population and is traditionally denoted by μ:

$$\mu = \sum_{i=1}^{N} X_i / N$$

For populations with different magnitudes, the location parameters must be different and therefore, as a single measure, the mean differs (Figure 2a). Knowing the mean “summarizes” the population. For many populations (but not all), the mean is a parameter.

So, how would the location parameter (the mean) of a population be estimated, given that the entire population cannot be measured? The mean of an unbiased, representative sample must be an accurate measure of the mean of the entire population. The only really representative samples are those that have the same frequency distribution as that of the population being sampled. If the sample reflects the population properly (i.e. has the same frequency distribution and is therefore representative), its location must be that of the population

For a population of N elements, from which a representative sample of *n* elements is measured:

$$\mu \approx \sum_{i=1}^{N} X_i / N$$

is the Location parameter (the mean) and

$$\bar{X} = \sum_{i=1}^{n} X_i / n$$

is an unbiased estimate of μ and is the sample mean.

Now it is clear what will happen if a sample is not representative of the population. If, for example. larger individuals in a population are more likely to be sampled, $\bar{X} > \mu$. The location parameter will be over-estimated because of the bias (inaccuracy) of the sample. Some unrepresentative samples may still, by chance alone, produce accurate, unbiased estimates of the mean of a population, but, in general, the mean of a sample will not estimate the mean of a population unless the sample is unbiased.

There are other measures of location used in environmental biology. Two are the median and the mode. The median is the “central” value of a variable. It is the magnitude that exceeds and is exceeded by the magnitudes of half the population. It is thus the symmetrical centre of a frequency distribution. The mode is the most frequent value of a variable in a population. This may or may not be the same as either the mean or the median depending on how other parameters in the population change the shape of the frequency distribution.

2.1.5 The Dispersion Parameter

The second most important parameter that dictates the shape or form of a frequency distribution is the dispersion parameter. This determines the degree to which the population is scattered or dispersed around its *central* location (or mean). To illustrate this, consider two populations with the same frequency distribution, except for dispersion (Figure 2b). The one with the larger dispersion parameter is more scattered. Practically, this means that measurements in a population with a large dispersion parameter will be much more variable than in a population that is less dispersed.

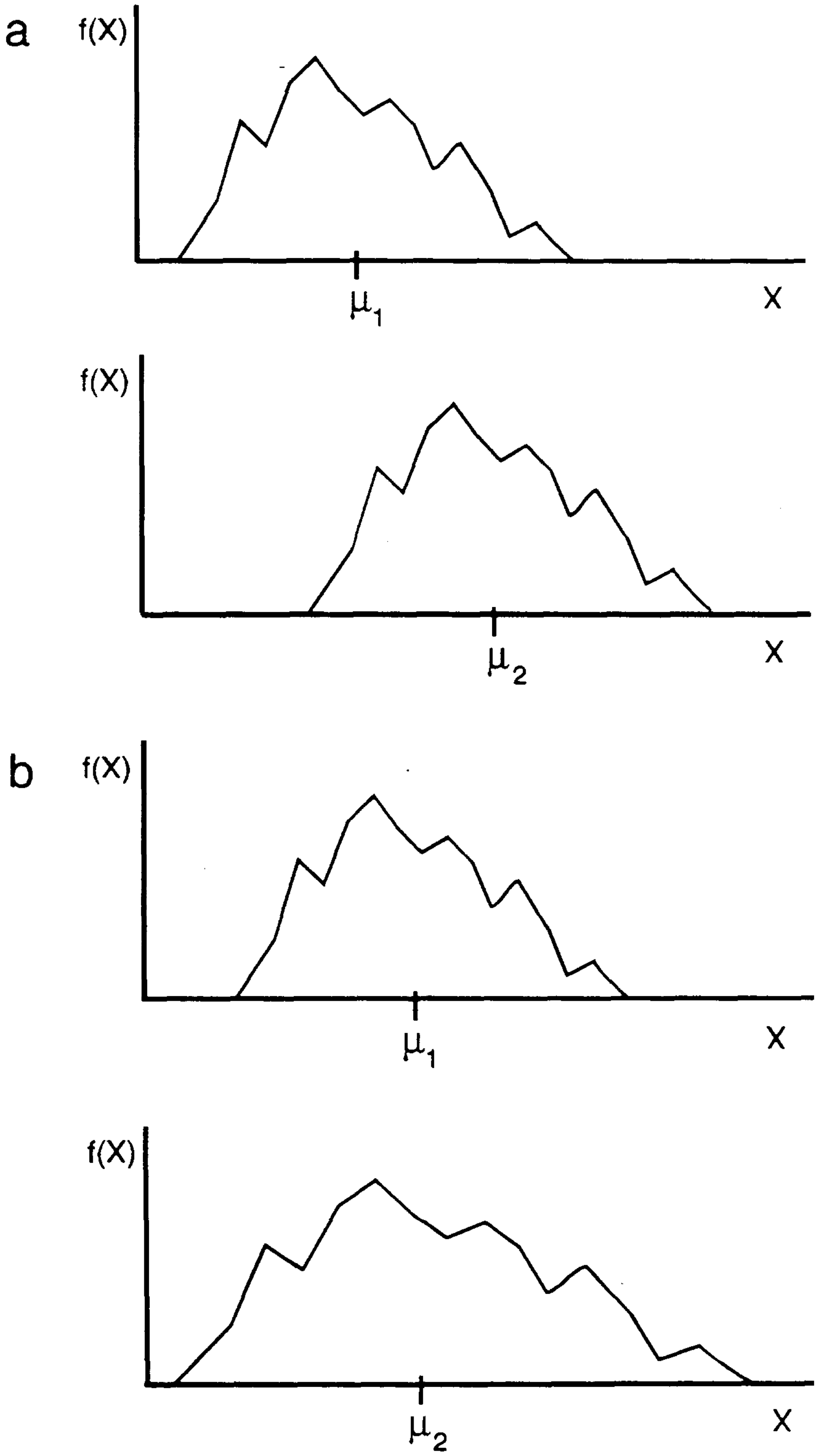

Figure 2. Effects of differences in location and dispersion parameters. (a) two populations of similar frequency distributions, but different location parameters (μ_2 is larger than μ_1). (b) two populations with similar frequency distributions and the same location parameter, but different dispersion parameters (σ_2^2 is larger than σ_1^2).

There are two reasons for needing information about the dispersion of a population. First, to describe a population, its dispersion is an important attribute. A much clearer picture of the population of interest can be gained from knowledge of its dispersion than from the simple description of its location. The second reason is much more important. An estimate of the dispersion parameter may be used to measure the precision of a sample estimate of the mean of a population. Thus, estimation of the population’s dispersion parameter will provide a measure of how close a measured, sample mean is to an unknown, population mean.

There are several possible measures of the dispersion parameter. For example, [he range of the population (smallest to largest value) is one measure. It is not, however, a practically useful one, because it is difficult to sample reliably. Only one member of the population has the largest value of any measurable quantity; another single member has the smallest value. They are not particularly likely to be contained in a sample (unless [he sample is very large compared to the whole population).

The most commonly used parameter for dispersion is the population’s variance, σ^2 It is preferred for several reasons, largely because its mathematical properties are known in terms of statistical theory. The arithmetic definition of a population’s variance is not intuitively interpretable. A logical choice of measure of dispersion would be a measure of how far each member of a population is from its location (i.e. its mean), In a widely dispersed population, these distances (or deviations) from the mean would be large compared to a less dispersed population.

A little thought makes it clear, however, that the average deviation from the mean cannot be useful because of the definition of location. The mean is the value which is, on average, zero deviation from all the members of a population. Thus, by definition of the mean, the average deviation from the mean must be zero.

If, however, these deviations are squared, they no longer add up to zero because the negative values no longer cancel the positive ones because all squared values are positive. Thus, the variance, σ^2, can be defined as the average squared deviation from the mean

$$\sigma^2 = \sum_{i=1}^{N} (X_i - \mu)^2 / N$$

The variance of a population is, however, in a different scale from the mean of the measurements - it is squared. To put the measure of variance in the same scale as the mean and the variable, we can use the Standard Deviation - the square root of the variance. σ is a measure of dispersion in the same scale as the variable being measured.

2,1.6 Sample Estimate of the Dispersion Parameter

As with the location parameter, common-sense suggests that the variance of a population can be estimated from a sample by use of the formula which defines the variance. This turns out not to be exactly correct and the appropriate formula for the sample estimate of variance, s^2, is:

$$s^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$

for a sample of *n* elements and for which the sample mean is $\overline{X}$

The estimates obtained from samples need to be described in terms of their *precision.* This is measurable in terms of the standard deviation (as explained earlier). It is. however, much more useful to see the standard error (see Table 1). This measures the variability of sample means. It takes into account the size of the sample (the number of replicates sampled) and is therefore smaller for larger samples. This is common-sense; where samples are larger, the mean of a sample should be more similar to the real mean being sampled than is the case where only small samples are available. So, standard error is a common description of precision of estimates. It combines information about the variance of the distribution being sampled and the size of the sample.

Table 1

Estimates of variance used to describe results of sampling and to indicate precision of samples. The calculations refer to a sample of *n* values of the variable X, Xl, X_2 .... X_i,.....X_n.

Statistical Estimate	symbol	Purpose
Variance	$s^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 / (n-1)$	Describes your estimate of the population's parameter, σ^2
Standard Deviation	s	Most useful description of the variability among replicate units in the sample, i.e. variation among the things counted or measured. Is in the same scale and units as the sample mean.
Standard Error	$s/\sqrt{n}$	Useful description of the variability to be expected among sample means, if the population were sampled again with the same size of sample. Is in the same scale and units as the sample mean.
Confidence Interval	$t_p \cdot s / \sqrt{n}$	Indicates [with a stated probability (1 - *p*)] the range in which the unknown mean (μ) of the population should occur. For a given size of sample and probability, specifies the precision of sampling. t_p comes from a table of the r-distribution with *(n* - 1) degrees of freedom, for a sample of *n* replicates. You choose the probability, *p*, commonly *p* = 0.05, in which case (1 - *p)* represents the 95% Confidence Interval.

A much better way of describing the values estimated from samples is to use confidence limits. Confidence limits are statements about the similarity between a sampled estimate of a mean and the true value being estimated. It is common to construct 95% confidence limits (as in Table 1) which are limits around an estimated mean that have a 95% chance of including the real value in the entire population. In other words, for any given sample, there is a 95% chance that the real mean is within the constructed limits. If the limits are small, the true value has a 95% chance of being close to the sampled value, so the sampled value is a precise estimate.

Whenever a sample is taken to estimate any parameter of a population of variables (e.g. the mean), the estimated value of the parameter, the size of the sample and a measure of variance or precision must be given (e.g. variance, standard deviation, standard *error* or confidence interval).

Note that there are many standard reference-books to explain sampling. Cochran and Cox (1957) is still one of the best. To increase precision of estimates, sampling can be designed to be much better able to reduce variability, For example, if the average number of plants per m^2 must be estimated over an extensive site, the variance of a random sample of quadrats will be very large. If, however, the area is divided into smaller regions, some of which appear to have many plants and some few plants, sampling in each region will provide more precise estimates for the averages in each region. This occurs because the large differences in numbers between regions do not also occur between replicates in the sample. The replicates are only scattered in each region and therefore have much more similar numbers of plants than occur across regions. Such stratified samples can then be combined to give an overall estimate of mean number of plants which will have a much smaller variance than would be the case from a random sample (Cochran 1957).

2.1.7 Representative Sampling and Accuracy of Samples

Note that the requirement to take a representative sample is not the same as the need to take a *random* sample. A random sample is one in which every member of the whole population has an equal and independent chance of being in the sample. Random samples will, on average, be representative. This is one reason why random samples are common and why procedures to acquire them are so important. Furthermore, random sampling is generally considered to be more likely to lead to independence of the data within a sample - an important topic covered later. The estimate of any parameter calculated from a random sample will be unbiased - which means that the average expected value of the estimates from numerous samples is exactly the parameter being estimated.

Why are simple random samples not always ideal? It is possible that a sample might provide an overestimate or underestimate of the mean density or the variance (Figure 3), because, by chance, the small, but randomly chosen number of quadrats all happened to land in patches of large (or small) density. If one knew that this had occurred, those quadrats would not be used (because we would know that they ‘were not representative). Even though we do not know what is the pattern of dispersion of the organisms is, we would normally assume that the animals or plants are not likely to be uniformly spread over the area. Much empirical ecological work has demonstrated the existence of gradients and patchiness in the distributions of animals and plants. If we had a map of densities, we would reject many ever-so-random, samples of quadrats because we could see they were not representative. The problem of non-representative samples is that estimates calculated from them are biased and they are inaccurate - they do not estimate correctly the underlying parameters of the distribution being measured.

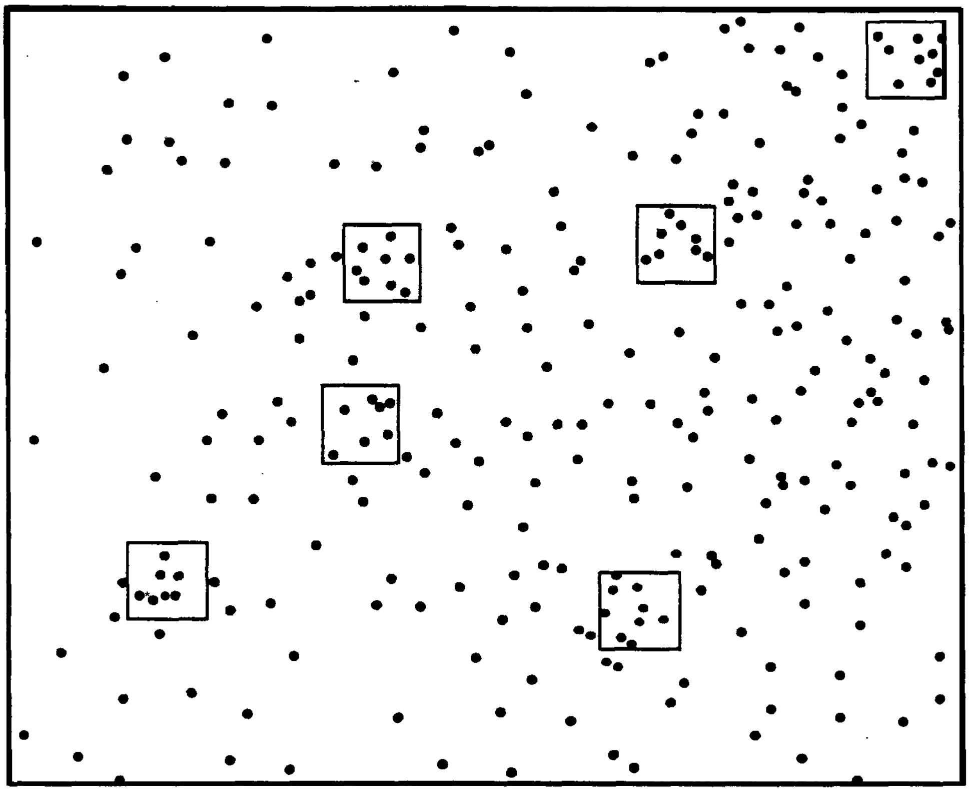

Figure 3. Random sampling with quadrats, placed in a field by randomly-chosen coordinates. By chance, the sample is unrepresentative because most of the quadrats have landed in patches of dense plants.

So, why is there so much emphasis on random sampling? The reason is that there are many sources of bias, including conscious and unconscious biases introduced by the observer or sampler. Random samples are not subject to some of these biases.

An example of unconscious (i.e. not deliberate) bias is selecting a sample of five fish from a tank containing 100 fish (the population being sampled to estimate mean rate of swimming under some different concentrations of pollution). It is possible (it may even be very likely) that the speed of swimming is inversely related to matchability in a small dip-net (i.e. the faster fish are harder to catch). If the fish are sampled by chasing and catching them one after another with a small dip-net, their average speed is going to underestimate that of the whole population, because you catch the five slowest fish! Here, the bias in sampling would lead to inaccurate (under) estimates of the mean speed. If, however, the fish were individually tagged with a number, or were all caught and placed in individually numbered containers, an unbiased sample could easily be obtained. Random numbers could be used to pick the individual fish (or containers), thus ensuring that the individuals sampled were chosen with no regard to their rate of swimming - which could then be measured objectively.

In cases when the whole population of individual entities to be sampled is available, you can use random numbers to pick representative samples very easily. You number the individuals (1 . . . . . *N)* and then pick the ones you want or need in your sample (1 . . . . *n)* by *n* different choices of random numbers in the range 1 . . . . *N,* i.e. discarding numbers that have already been chosen (otherwise those individuals have a greater chance of appearing in the sample). If the population to be sampled is very large, it may prove impracticable to number all the individuals. Some form of “two-stage” sampling is the answer, The specimens could be divided into arbitrarily sized groups (of say 50 each). Then *n* groups would be chosen at random and one organism randomly chosen out of the 50 in each sampled group. An unbiased sample will result if every organism in a group has an equal chance of being picked and every group of 50 organisms has an equal chance of being picked.

This can also be done in the field. A sample could probably best be chosen by picking *n* randomly-placed small quadrats (or other sampling unit) in the field. In each quadrat, the organisms can be numbered and an individual chosen at random,

These are relatively straightforward procedures. There are, however, traps in attempting to find representative samples using random numbers. Consider trying to pick a sample of quadrats in an area so that the density of organisms per unit area can be estimated for the entire area. There are two commonly used procedures. In one, the population of quadrats is sampled by choosing *n* quadrats at random (Figure 4a).

The alternative method consists of finding a random position for the bottom left corner of a quadrat to be sampled. This does not lead to an unbiased sample, as illustrated in Figure 4b. There is less chance of sampling some parts of the area than everywhere else. Position A has more chance of being included in a sampled quadrat than is the case for points B and C, because more points can be chosen to be the bottom left corner of a quadrat containing A than of a quadrat containing B. No quadrat can start outside the left hand edge of the area to be sampled and no quadrat can start so far to the right that it extends over the right-hand edge of the area to be sampled.

3. DOING STATISTICAL TESTS

3.1 MAKING HYPOTHESES EXPLICIT

We proceed in ecology or environmental science or any other science, by making observations about nature and then attempting to explain them by proposing theories or models. Usually, several possible models will equally well explain some set of observations. So, we need some discriminatory procedure to distinguish among alternative models (or theories). Therefore, we deduce from each model a specific hypothesis (or set of hypotheses) which predicts events in some as-yet-undocumented scenario, if the model is correct. If a model is incorrect, its predictions will not come true and it will therefore fail - provided that it is tested. Experiments , are therefore tests of hypotheses. An experiment is the test that occurs when the circumstances specified in the hypothesis are created, so that the validity of the predictions can be examined. Sampling in a different area or under a different set of conditions is an experiment as defined here - as long as the model being evaluated and hypothesis being tested are explicit and the sampling program programme has been designed to address these Because of the logical nature of “proof” (see Hocutt 1979, Lemmon 1991), we must usually attempt to turn an hypothesis into its opposite (the “null” hypothesis) and then do the experiment (or sampling) in an attempt to disprove the null hypothesis. This will provide empirical and logical support for the hypothesis and model but will never prove that any model or theory is correct. This cannot be done.

Statistical procedures are usually appropriate to help decide whether or not to rejector to retain the stated null hypothesis. This causes two very different problems for the environmental scientist. First, statistical null hypotheses are often quite different from logical null hypotheses, causing immense potential for confusion and quite invalid (i.e. illogical) inferences in certain types of sampling procedures. The details of these problems are discussed in Underwood (1990, 1994a). The second issue is that" statistical analyses used to help make decisions about rejection or retention of a null hypothesis almost invariably require assumptions - often quite strict assumptions - about the data gathered during an experiment.

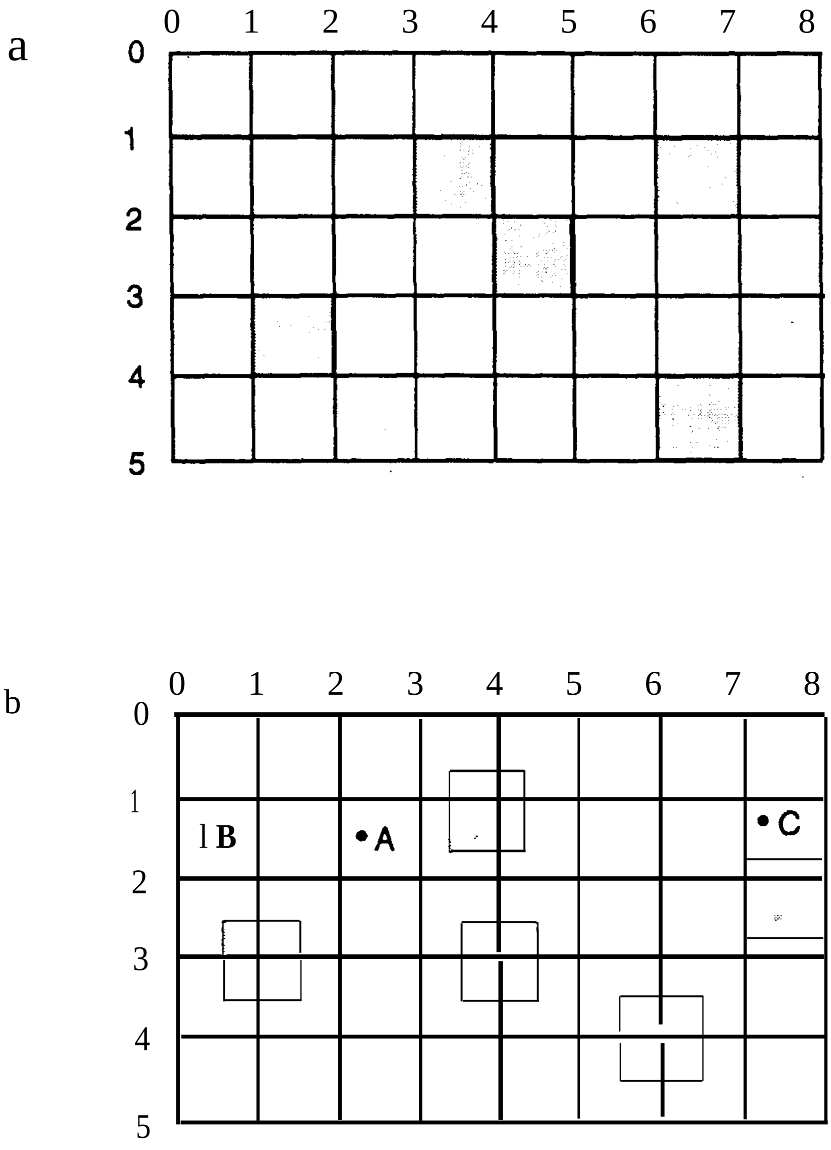

Figure 4. Random (representative) sampling of a specified study-site of sides 8 x 5 m. Quadrats (1 x 1 m) are placed by coordinates measured from the top left-hand comer. (a) Five (shaded) quadrats of a total of 40 are chosen by random coordinates at 1 m spacing. (b) Quadrats are sited at 0.1 m intervals. Points A, B and C are discussed in the text.

It is crucial for any experiment that the hypothesis (or hypotheses) be identified wisely and well. This will be illustrated here by an example of environmental sampling to determine the success of a programme of rehabilitation of a polluted site after mining. Although this is an example from a terrestrial system, the same principles will apply to marine systems after, say, dredging or removal of mangrove forests At the start, there are few plants in the area mined because the mine removed them. Work is done to replant the appropriate terrestrial vegetation so that the original biomass or cover or diversity of species will be recreated. If sufficient biomass, cover or species are later found in the area, the site will be declared rehabilitated. The usual approach to this is based on the original biomass of plants in the mined site being different from that in a control site (or better, a series of control sites). The typical hypothesis can therefore be stated as “the ratio of the average biomass of plants in the disturbed site to that in the control(s) will be less than one”. The null hypothesis is that the ratio of mean biomasses will be 1 (or, in fact, the mined site will have more weight of plant material per unit area than occurs in controls and the ratio will be greater than 1).

When the plants in the previously mined and the control sites are sampled, the hypothesis of no difference can be tested. If the null hypothesis continues to be rejected in statistical tests, the mined site cannot be declared rehabilitated and the program of remediation must continue. This classical view leads to a very serious, if not environmentally disastrous problem.

If poor sampling is done, leading to very imprecise estimates of the mean values of the variable being sampled, it will be easy to demonstrate that there is no difference between the control and disturbed sites, even if they are very different. This is illustrated in Figure 5a. In contrast, if the mean value of the chosen variable really is very similar between the disturbed and control sites, but sampling is very precise, the statistical test will keep rejecting the null hypothesis (Figure 5b). Under these circumstances, expensive remediation will have to be continued, even though there is no persistent loss of plant biomass.

Obviously, there will be considerable financial pressure on those responsible for rehabilitation to use sloppy and imprecise methods of sampling. The problem is the defined hypothesis and null hypothesis. What is needed is a reversal of the hypothesis (McDonald and Erickson 1994). This could only be accomplished by defining what minimal mean biomass of plants would be acceptable as an indication that the site is minimally recovered. This is illustrated in Figure 5c, where it was decided that if the biomass of plants was about three-quarters of that in an undisturbed, control area, this would represent “recovery”. Once this has been defined. it is then only possible to use precise sampling to demonstrate equivalence of abundance of biomass in the two areas, as shown in the lower parts of Figure 5c. If imprecise estimates are now obtained from poorly designed sampling, the data will not cause rejection of the null hypothesis. The disturbed site will continue to be declared as not rehabilitated. Only by having appropriate precise estimates will it be possible to demonstrate that the minimal conditions required to demonstrate rehabilitation have been met.

This example serves notice that considerable care must be put into the appropriate methods and logics of environmental sampling. Only where appropriate hypotheses are clearly defined can sensible interpretations be made.

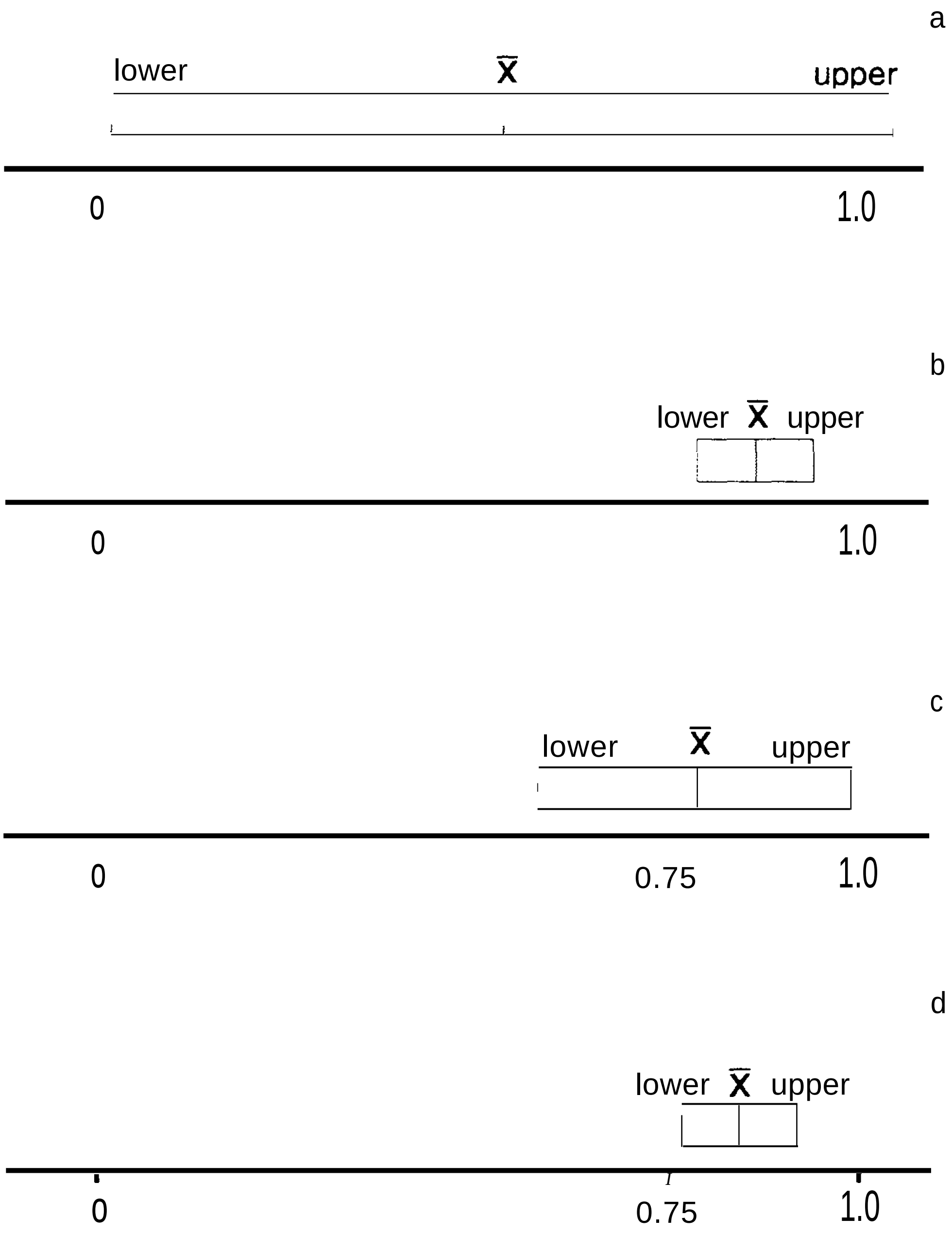

Figure 5. Assessment of recovery of a disturbed site. In all four diagrams, the ratio of biomass of plants in a previously mined area to that in a control area is shown as the theoretical value of 1, which represents no difference, i.e. complete recovery of the population in the previously mined area. In each case, a sample is taken, with mean abundance of plants, $\bar{X}$, and upper and lower boundaries of confidence limit, as shown. In (a) and (c), sampling is sloppy, precision is not great and confidence limits are quite large. In (b) and (d), sampling is precise and confidence limits are small. Under a traditional null hypothesis (the disturbed and control sites do not differ), situation (a) would erroneously retain the null hypothesis - solely because of sloppy sampling. The two sites are clearly different. In contrast, under the traditional null hypothesis, situation (b) would cause rejection of the null hypothesis, even though the two sites are really similar. This is because the sampling is so precise that small differences seem to be important. In (c) and (d), the null hypothesis is that the ratio of mean numbers between the two sites is less than 0.75- the minimal ratio being considered to represent recovery of the population (this was arbitrarily chosen for this example). Now, sloppy sampling in (c) leads to retaining the null hypothesis (as it should). Precise sampling allows rejection of the null hypothesises and therefore a correct

declaration that the two sites are similar and recovery is satisfactory (after McDonald and Erickson 1994).

3.2 THE COMPONENTS OF A STATISTICAL TEST

3.2.1 A Null Frequency Distribution

First, the null hypothesis must be defined in some quantitative manner. This turns a logical null hypothesis into a statistical null hypothesis - that defines a frequency distribution that can be sampled to examine the likely validity of the logical null hypothesis. The statistical null hypothesis therefore defines the frequency distribution of possible results,

3.2.2 A Test Statistic

Then, a test statistic is chosen. This has two properties. It must be measurable in some possible experiment or sampling protocol. Second, it must be possible to calculate in advance the frequency distribution of the test statistic if the null hypothesis is true.

Now a decision must be made about how unlikely an event must be (if the null hypothesis is true) so that we would consider the event more likely to represent an alternative hypothesis. The conventional probability is 0.05 (1 in 20) but this is entirely arbitrary. This defines the region of rejection of the null hypothesis. We would reject the null hypothesis if the experiment or sampling program produces a result which has less than a probability of 0.05 of occurring by chance if the null hypothesis were true. So, the region of rejection of the null hypothesis is all parts of the frequency distribution of the test statistic which would cause us to reject the null hypothesis. The boundaries of the region of rejection are defined by the Critical Values of the experiment. If we get a result equal to or greater than the upper critical value or equal to or smaller than the lower critical value we will chose to reject the null hypothesis.

All statistical tests have the same component steps. You must define the critical values before the experiment or sampling is done. Otherwise, there is no way of interpreting its outcome. For example, if the region of rejection and therefore the critical values are chosen after the experimental result is known, they could be chosen to be outside or inside the observed value of the test statistic. Choosing to place the critical values outside the observed test statistic means the latter does not exceed the former and the null hypothesis will be retained. Putting the critical values inside the observed value means the latter is outside the former and therefore in the region of rejection, causing the null hypothesis to be rejected. By choosing where to put the critical values (by defining the region of rejection) after the experiment, the procedure by which to reach a conclusion from the experiment could be altered to produce whatever conclusion is chosen by the experimenter. This obviously makes no sense in the construction of a tool to aid in making decisions. If the decision is made without use of the procedure (which it would be if it were decided whereto place the region of rejection after knowing the outcome of placing it either side of the observed value) then the procedure and the experiment were not necessary.

The correct procedure is to identify how a choice would be made to retain or reject the null hypothesis before the data are obtained. Then. stick to the decision once the data cause it to be made. Later, these will be examples of statistical tests and how a decision can be made using one. In principle, the test statistic's frequency is known and the lower and upper critical values are defined. This is shown in Figure 6, which is the distribution of a statistic called *t*, widely used in statistics to test null hypotheses such as:

$$H_0 : \mu = h$$

where μ is the mean value of some population and *h* is some specified number. For example, if it is suggested that the mean weight of livers of fish in an area is 23 g, we could test this idea by proposing the null hypothesis:

$$H_0 : \mu_{weight} = 23$$

t is calculated from the sample as:

$$t = (\overline{X} - 23) / \text{S.E.}$$

where $\overline{X}$ is a sampled estimate of the mean and S.E. is the standard error of the sample. If the null hypothesis is true, *t* should be centred on 0 (all samples should have means scattered around 23). If hundreds of samples were taken, their values of *t* would be the distribution in Figure 6. Only 2.5% of all the values would be larger than the upper critical value. Only 2.5% of all the values would be smaller than the lower critical value.

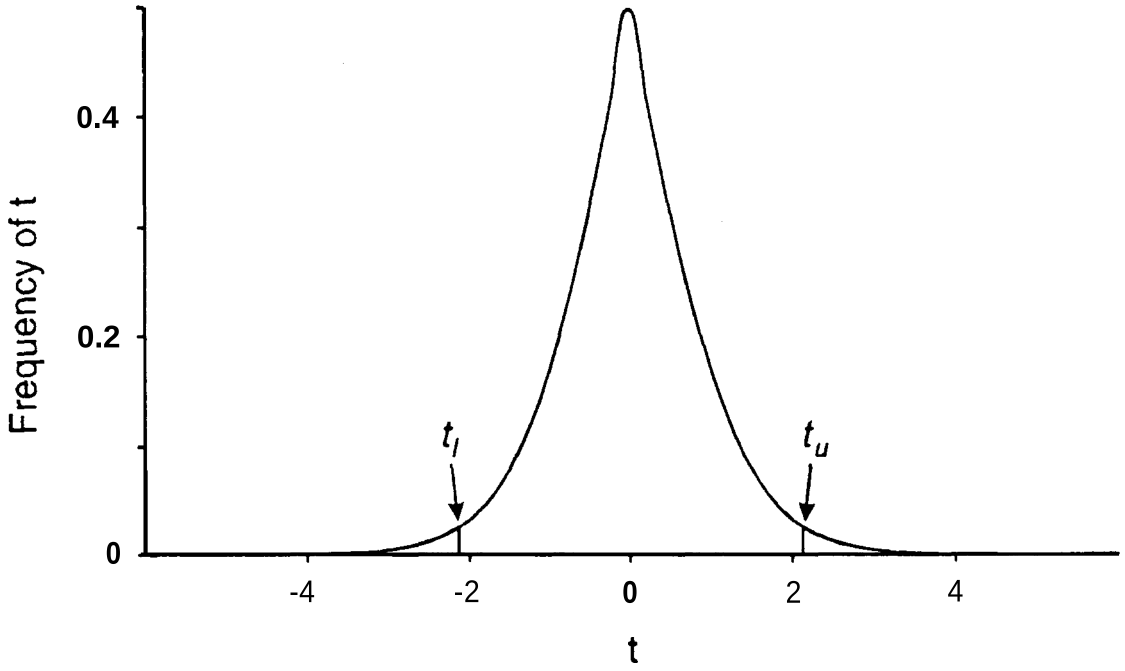

Figure 6. Frequency distribution of Student's t-statistic for samples of size n = 15, i.e. 14 degrees of freedom. t_u and t_l are the upper and lower critical values, each cutting off 2.5 % of the distribution.

If we take a sample and obtain a value oft between the lower and upper critical values, our sampled value of *t* is consistent with 95% of the possible values. It is therefore consistent with the idea that the mean is 23. On the other hand, if our observed value of *t* was larger than the upper critical value (t_u in Figure 6), we should reject the null hypothesis in favour of the alternative that μ > 23. Such large values of t only recur in 2.5% of all possible samples when the null hypothesis is true, but are quite likely if the true mean is greater than 23. In the same way, if the observed value of t from our sample was smaller than the lower critical value (t_l in Figure 6), we should conclude that the mean weight was smaller than 23.

3.2.3 Type I and Type II Errors in Relation to a Null Hypothesis

Table 2 shows the only four possible outcomes of a statistical test. If the null hypothesis is true and is retained as a result of the statistical test or is false and is rejected, the correct inference has been made. If the null hypothesis is true, but is rejected, the experimenter has made a mistake. The probability of making such a mistake is, however, entirely under the control of the experimenter. It is chosen, in advance of the statistical test by the choice of critical values and the region of rejection. So far, this has been arbitrarily determined to be a probability of 0.05, one in twenty. This probability of Type I error is commonly known as the probability *a*. In the case of the t -test described above, a = 0.05 or 5%.

Table 2

Type I and Type II errors in interpretations of results of experiments.

	Null Hypothesis (H_0) is (unknown to us):	
Outcome of statistical test is to:	TRUE	FALSE
REJECT H_0	Type I error occurs with probability (a) chosen by experimenter	Correct conclusion: False H_0 is rejected
RETAIN H_0	Correct conclusion: True H_0 is retained	Type II error occurs with probability (β) sometimes chosen by experimenter

If the null hypothesis is true, 2.5 % of all the possible samples will give values of t smaller than the lower critical value and will therefore cause us to reject the null hypothesis even though it is true. Another 2.5 % of samples would cause r to be larger than the upper critical value. So, there is a 5 % chance of getting a sample that causes rejection of the null hypothesis when it is true.

Rejection of a true null hypothesis occurs whenever a sample estimate of the parameters of the population generates a test statistic in the region of rejection. We keep a small so that such samples are unlikely.

There is, however, the opposite error which occurs when the null hypothesis should be rejected because it is wrong. A sample, however, may cause calculation of a test statistic that is consistent with the null hypothesis. How this comes about is illustrated in Figure 7. The first distribution is the normal distribution of a variable, centred on the mean specified by the null hypothesis (μ = 23), The second distribution, is the distribution of the variable (with the same variance as for the previous distribution) with a mean of 25.3, some 10 % larger than the one specified in the null hypothesis.

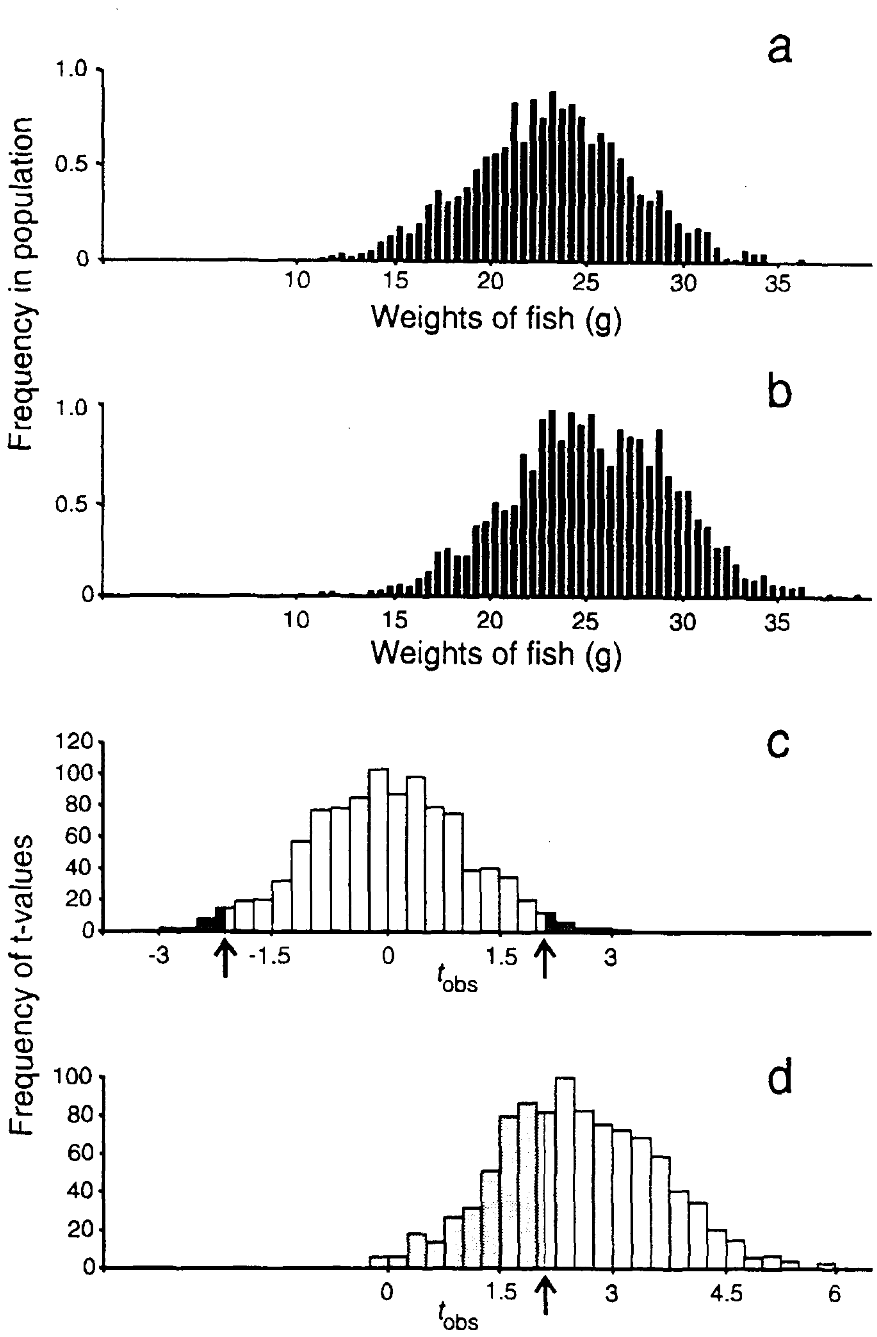

Figure 7. Demonstration of Type I and Type II error in t-tests. The mean weight of fish is sampled to determine whether they weigh, as claimed, 23 g (i.e. the null hypothesis is H_0: μ = 23). The weights are normally distributed. In (a) and (b), the distributions of weights of fish are shown when the null hypothesis is true and μ = 23 (a) and when H_0 is false and the mean is larger, μ = 25.3 (b). In (c) are observed values of t_{obs} from numerous samples of population (a), with $n = 20$. The probability of incorrectly rejecting the null hypothesis (α, the probability of Type I error) is chosen to be 0.05. The arrow indicates the critical value of r at $P = 0.05$, $t_{crit} = 2.09$, with 19 degrees of freedom. The shaded areas ($t > 2.09$ and $t < -2.09$) show this probability. In (d) are the observed values of ^{t}obs from numerous samples of population (b). The probability of incorrectly retaining the null hypothesis (β, the probability of Type II error) is shown as the dotted area, i.e. all values in distribution (d) that are smaller than 2.09, the upper critical value for distribution (c).

Below these (Figure 7c) is the null distribution oft - values; this is the distribution of values of t that would be obtained from repeated independent sampling of the distribution of the variable when the null hypothesis is true. Finally, there is the distribution of values of t that would be obtained from repeated sampling of the alternative distribution, Note that the two distributions of t overlap. Thus, some samples obtained from either the null or the alternative distribution of weights of livers of fish will generate values of t that could come from either of the two frequency distributions of t.

The region of rejection was specified in advance as a = 0.05 (i.e. the probability of Type I error is 0.05), as shown by the shaded area in Figure 7c. A value of t_{obs} of, say, 1.5 would therefore not cause rejection of the null hypothesis. [f the null hypothesis were true, this value of t would have come from the null distribution and the conclusion would be correct, On the other hand, if the null hypothesis were false, this value of t_{obs} would have come from the alternative distribution and the conclusion would be incorrect - a Type II error,

4. INTRODUCTION TO ANALYSIS OF VARIANCE

One of the most powerful and versatile tools in experimental design is the set of procedures known an analysis of variance. These start from a simple conceptual basis, but can be used in many different aspects of biological research. They are very suitable for planned, manipulative experiments and are also for environmental sampling.

The hypothesis underlying all analyses of variance is that some difference is predicted to occur among the means of a set of populations. The null hypothesis is always of the form:

$$H_0\colon \mu_1 = \mu_2 \ldots\ldots = \mu_i \ldots\ldots = \mu_a \; (= \mu)$$

where μ_i represents the mean (location) of population i in a set of a populations. It is proposed as the null case that all populations have the same mean μ. Experiments with such a linear array of means are known as single factor or one-factor experiments. Each population represents a single treatment or level of the factor.

Any departure from this null hypothesis is a valid alternative - ranging from one population being different from the other $(a$ - 1) populations to all a populations differing. Departure from the null hypothesis does not imply that all the populations differ.

4.1 ONE-FACTOR ANALYSIS OF VARIANCE

To test the null hypothesis given above, a representative sample is taken of each of the populations. In the simplest case, these samples are balanced - they are all of the same size, n.

The total variability among the numbers in the entire set of data can be measured by calculating how far the individual values (the X_{ij}'s) are from the overall mean of all data combined $\overline{X}$. The deviations are all squared. The total variation amongst the entire set of data (i.e. the sum of squared deviations from the mean over the entire set of data) is known as the total sum of squares.

$$\text{Total Sum of Squares} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}\right)^2$$

The total Sum of Squares can be divided into two components (see , for example, Underwood 1981):

$$\text{Total Sum of Squares} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_i\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(\bar{X}_i-\bar{X}\right)^2$$

The first of these two terms measures some function of variability **within** the samples - it is calculated as deviations of the data (X_{ij}'s) from the mean of the sample to which they belong ($\bar{X}_i$ ‘s). The second term obviously involves variability **among** the samples - it is calculated as the deviations of each sample mean from the overall mean. The two sources of variability (Within and Among Samples) add up to the total variability amongst all the data When the null hypothesis is true, the sample means should all be the same except for sampling error.

4.2 A LINEAR MODEL

Suppose you intend to count the number of worms in replicate cores in several (u) places in an estuary. In each place, there is a frequency distribution of number of worms per core. In place i, this distribution has given number μ_i and you take a sample of n cores (1..... j...... n). In core j, in sample i, the number of worms is X_{ij}.

The model developed to describe data from such a sampling programme is:

$$X_{ij} = \mu_i + e_{ij}$$

where μ_i is the mean of the population from which a sample is taken and X_{ij} is the jth replicate in the ith sample. e_{ij} is therefore a function of the variance of the population - it measures how far each X_{ij} is from the mean of its population. This is illustrated diagrammatically in Figure 8. The X_{ij}'s are distributed with variance σ_i^2 around their mean μ_i

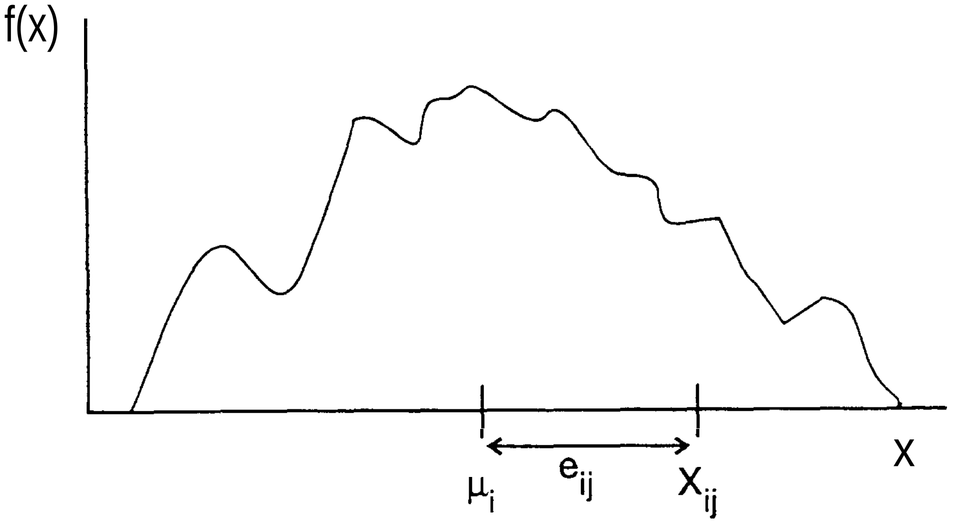

Figure 8. Diagram illustrating the error term in an analysis of variance. Frequency distribution of a variable X, with mean μ_i and variance σ_i^2 X_{ij} is a particular value, at a distance e_{ij} from μ_i

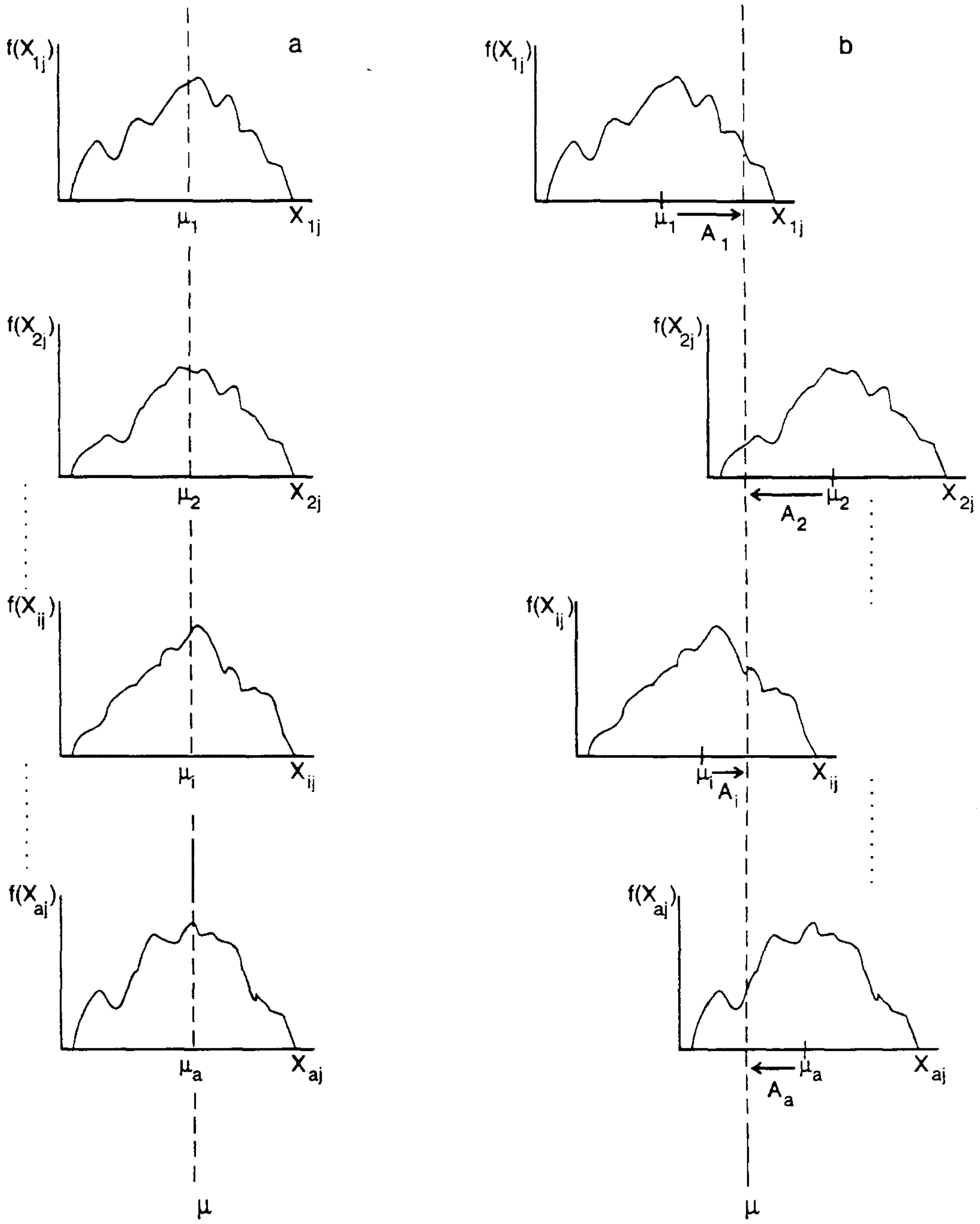

Figure 9. Illustration of A_i terms in the linear model for an analysis of variance. There are 1 ... i . . . . . . a populations of the variable X_{ij}, each with mean, μ_i . In (a), all populations have the same mean, μ ; the null hypothesis is true. In (b), the null hypothesis is false and some populations have different means, which differ from p., the overall mean, by amounts A_i.

In terms of the null hypothesis, all the populations have the same mean μ . If the null hypothesis is not true, the populations do not all have the same mean and it can be assumed that they differ (at least, those that do differ) by amounts identified as A_i terms in Figure 9.

$$\therefore X_{ij} = \mu + A_i + e_{ij}$$

where μ is the mean of all populations, A_i *is* the linear difference between the mean of population i and the mean of all populations, if the null hypothesis is not true and e_{ij} is as before. Within any population, a particular individual value (X_{ij}) differs from the mean of its population by an individual amount (e_{ij}) When the null hypothesis is true, all values of A_i are zero (see Figure 8).

The two sources of variation can be interpreted by substitution of the terms by their equivalent expressions from the linear model. The results are shown in Table 3. The derivation of these is given in most text-books (see Snedecor and Cochran 1989; Underwood 1981; Wirier et al. 1991).

These results are only true if the populations are sampled independently (see later for what this means). So, the first assumption we must make about the data is that they are independent samples of the populations. Also, the data must come from populations that have the same variance (and, but less importantly, the populations should have normal distributions). We must also assume that the populations all have [he same variance:

$$\sigma_1^2 = \sigma_2^2 \ldots\ldots = \sigma_i^2 \ldots .= \sigma_a^2$$

If this assumption is true, it is equivalent to stating that all populations have the same variance, σ_e^2

4.3 DEGREES OF FREEDOM

The next thing needed is the degrees of freedom associated with each sum of squares. For the Total Sum of Squares, there is a total of n data in each of the a samples and therefore an data in the set. The degrees of freedom are an -1. For the variation among samples, there are a sample means ($\overline{X}_i$ ‘s) and $(a$ - 1) degrees of freedom. There are $a(n$ - 1) degrees of freedom within samples.

4.4 MEAN SQUARES AND TEST STATISTIC

The final manipulation of the algebra is to divide the sums of squares for the two sources of variation by their degrees of freedom to produce terms known as Mean Squares, as in Table 3.

Table 3

Analysis of variance with one factor. Mean squares estimate terms from the linear model

Source of Variation	Sums of Squares (SS)	Df	Mean Square	Mean Square Estimates
Among Samples	$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(\overline{X}_i-\overline{X}\right)^2$	$(a$ -1)	$\dfrac{\sum^{a}\sum^{n}\left(\overline{X}_i-\overline{X}\right)^2}{(a-1)}$	$\sigma_e^2+\dfrac{\sum^{a}\sum^{n}\left(A_i-\overline{A}\right)^2}{(a-1)}$
Within Samples	$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(X_{ij}-\overline{X}_i\right)^2$	$a(n$ - 1)	$\dfrac{\sum^{a}\sum^{n}\left(X_{ij}-\overline{X}_i\right)^2}{\sum^{a}(n_i-1)}$	σ_e^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(X_{ij}-\overline{X}\right)^2$	an -1		

When the null hypothesis is true, the A_i terms are all zero, so the Mean Square among samples estimates only the variance of the populations σ_e^2 and so does the Mean Square within samples (see Table 3). Therefore, MS_{Among} / MS_{Within} estimates 1 (except for sampling error).

Because the Mean Squares are composed only of squared numbers (i.e. the $(A_i - \bar{A})^2$ cannot be negative), the only possible alternative to the null hypothesis is:

$$H_A: \frac{MS_{Among}}{MS_{Within}} > 1$$

The distribution of the ratio of two estimates of a variance when the null hypothesis that they are equal is true is known as F-ratio. It is therefore a convenient test statistic for the present null hypothesis

Having derived the formulae for calculating sums of squares as measures of variation among data, they should not be used! These formulae involve considerable rounding errors in their numeric results, because means are usually rounded somewhere in the calculations. Instead, algebraically identical formulae which introduce little rounding should be used. These are given in numerous text-books and commercial software.

If the region of rejection of the test-statistic is chosen by having the probability of Type I error $\alpha = 0.05$, there is only one test to evaluate all potential differences among means. A worked example is given in Table 4

Table 4

Data for 1-factor analysis of variance

a) Example of data; *a* = 4 samples, n = 5 replicates

	Treatment			
Replicate	1	2	3	4
1	414	353	50	413
2	398	357	76	349
3	419	359	28	383
4	415	396	29	364
5	387	395	27	365
Sample mean	406.6	372	42	374.8
Sample variance	184.3	465	452.5	601.2
Standard error	6.07	9.64	9.51	10.97

Overall mean = 298.85

b) Analysis of variance

Source of variation	Sums of Square	Degrees of freedom	Mean Square	F-ratio	Probability
Among samples	34463.6	3	11487.9	18.82	<0.0001
Within samples	9768.4	16	610.5		
Total	44237.0	19			

4.5 UNBALANCED DATA

For the single-factor analysis, there is no need for the sizes of samples to be equal or balanced. Before taking unbalanced samples, there ought to be some thought about why it is appropriate. The null hypothesis is about a set of means. Why some of them should be sampled more precisely than others, which will be the case for those with large sample sizes, cannot usually be justified.

Sometimes. however, samples become unbalanced due to loss of data, death of animals, arrival of washed-up whales rotting the plot, plane-crashes, human interference, There are several ways to proceed in order to complete the analysis.

If there are few replicates in nearly every treatment and only the odd one is missing here or there, do the analysis anyway. Use all the data you have because you have very few data! If there are numerous replicates (say. $n \geq 20$) and only an occasional one is lost, do the unbalanced analysis because this amount of difference is small and the effect on precision is minuscule when n is large.

What should be avoided is the situation where some samples have large and others have small n. This leads to great imbalance in the precision of sampled means. More importantly, it leads to different power to detect differences among samples when multiple comparisons are done. If five treatments have n = 5, 5, 10, 10, 20. respectively, there will be greater precision in a comparison of treatments 3 and 4 with 5 than a comparison of treatments 1 and 2 with 3 or 4.

It is usually simpler, in addition to being better, to use balanced samples

4.6 ASSUMPTIONS OF ANALYSIS OF VARIANCE

There are several assumptions underlying all analyses of variance,

(i) Independence of data within and among samples;

(ii) Homogeneity of variances for all samples;

(iii) Normality of data in each distribution.

4.6.1 Independence of Data

This problem - ensuring that data are independently sampled - is so great that many environmental and ecological studies are invalidated by failing to solve it. This has, unfortunately, not prevented the publication of results of such studies, As a result, subsequent sampling and experiments are done using published methods and designs even though these are wrong.

The effects of non-independence of data on interpretation of the results of analyses of variance are summarised in Table 5. Each of these can be avoided or managed, provided attention is paid to the necessity to remove them from experimental designs.

Table 5

Consequences for interpretation of experiments of non-independence among replicates (within treatments) or among treatments.

	Non-independence Within Treatments	**Non-independence Among Treatments**
Positive Correlation		
	Variance within samples under-estimated	Variance among samples under-estimated
	F-ratios excessive	F-ratios too small
	Increased Type I error	Increased Type II error
	Spurious differences detected	Real differences not detected
Negative Correlation		
	Variance within samples over-estimated	Variance among samples over-estimated
	F-ratios too small	F-ratios excessive
	Increased Type II error	Increased Type I error
	Real differences not detected	Spurious differences detected

The onus of creating independence always falls on the experimenter. Within limits set by the needs of the stated hypotheses, independent data can almost always be obtained by application of more thought and more effort and, occasionally, more money. It involves considerable thought about and knowledge of the biology (or chemistry, etc.) of the system being investigated.

There are, essentially, four types of non-independence (see the detailed discussion in Underwood 1994).

Positive Correlation Within Samples

You may be testing the effects of a particular pollutant on the growth-rate of a species - say, lobsters. Unknown to you, there may be a behaviourally-induced form of non-independence affecting these rates of growth. If the. lobsters are not influenced by the presence of other lobsters they are completely independent of one another.

In this case it makes no difference whether they are kept together in one aquarium or are kept separately in 15 different tanks. Now consider the result of sampling a group of lobsters which are influenced by the behaviour of other animals. Suppose the lobsters behave as described above when they are isolated in addition to initiating their own bouts of foraging, they are more likely to emerge to feed when another lobster is feeding. The total time spent feeding will be greatly increased. There will also be a substantial decrease in the variance of these variables within samples whenever there is a positive correlation among the replicates in a sample,

The consequences for statistical tests on a sample are very drastic. In this case, the increased mean time spent feeding (and therefore, probably, the rate of growth of the animals) was increased and any comparison of such factors on rate of growth would be under the conditions caused by non-independent behaviour. Only if any factors on rate of growth would be under the conditions caused by non-independent behaviour. Only if any differences were evident at enhanced rates of growth would they be detected by experiments. More generally, however, the influence of positive correlation among replicates is to cause excessive Type I error in statistical tests on differences among samples.

The positive correlation among replicates decreases the variation within samples. The mean square among treatments continues to measure the intrinsic variability among lobsters as though they were behaving independently because the estimate of variability among treatments comes from the means of each sample and the relationships between the variance of sample means and the variance of the distributions being sampled.

The result is dramatic. Because the estimated variation within samples is much smaller than that found when lobsters are independently sampled, the variability among samples is relatively large (even though it is correctly estimated). The resulting *F*- ratios are much larger than they should be if data were independently sampled. The distribution of *F is* markedly shifted to larger values and there is, consequently, massively excessive probability of Type I error.

If it is important to keep animals, or plots or whatever experimental units, separate because the null hypothesis requires it, keeping them together will almost certainly create some form of non-independence. Data will only be independently sampled by having independent replicates. On the other hand, if the null hypothesis is about groups of animals, or experimental plots or clumps of plants, etc., it is essential to have measures of the average results in replicated groups or set of plots or clumps - not from individual members of groups or clumps. This always results in more work because data from individuals cannot be used unless it can be demonstrated that they are independent.

Experimental biologists must always be on their guard against non-independence. The requirement for independent data is not a unique property of analysis of variance, nor of parametric versus distribution- free tests. It is a necessity due to underlying probability. It is, however, a biological problem. Biological (behavioral and ecological) processes cause non-independence among replicates.

Negative Correlation Within Samples

Non-independently sampled data may also be negatively correlated. Reconsider the previous example of lobsters, but this time , many will not feed if other lobsters are already feeding. The total time spent feeding by some lobsters will be very dramatically reduced compared to what happens if they are independently sampled (kept on their own). Of much more consequence to statistical tests, the sampled variances of each of these variables will be very much *larger* than for independent lobsters.

F- ratios testing for differences among samples are on average, much smaller than for independently sampled lobsters, because the variance among replicates is inflated. This reduces the probabilities of Type I error but inflates the probability of Type II error (retaining incorrect null hypotheses). In other words, if there were differences due to pollutants, they would have to be much larger to be detected in experiments with negatively correlated lobsters than with independently sampled ones. The **power** of the experiment to detect differences is dramatically reduced when replicates are negatively non-independent.

Negative Correlation Among Samples

There can also be negative correlations among experimental treatments. Suppose that two species of sessile animals, such as sponges, have been observed to have very patchy distributions at a small spatial scale. Patches of either species are about 5 to 50 cm in diameter. The hypothesis being tested is that one of the species is more abundant (i.e. occurs with greater percentage cover than the other. These measurements are made using quadrats.

There are two ways to proceed. In one, a sample of quadrats is thrown representatively in the study-area and both species are counted in each quadrat. These data are not independent. In the other, a sample of quadrats is thrown and the cover of Species A recorded. Then a second set of independently-placed quadrats is sampled to estimate the cover of Species B. These data are independent.

If the two species are sampled in the same quadrats, there is obviously going to be extreme negative correlation between them. Where one covers a lot of space in a quadrat, the other can only have a small cover and vice versa. This tends to maximize the difference in cover between the two species per quadrat. If, by chance, several quadrats happen to have above average covers of Species A, the mean cover will be over-estimated, but the mean of Species B must be under-estimated because the two estimates are tied together by the method of sampling. *F*- ratios will be larger than expected, causing a much increased probability of Type I error.

The only satisfactory solution to the problem of non-independence among experimental treatments is to do the extra work necessary to ensure the availability of independent sampling.

Positive Correlation Among Samples

The final form of non-independence that is of major concern to biologists is positive correlation among treatments. This will be illustrated by an experimental manipulation of prey of some predatory species that operates in a very patchy habitat.

This can occur when field experiments are done in very patchy habitats, but the replicates are not randomly dispersed across the patches. If the experimental treatments (e.g. a control and an experimental plot) are placed close together, so that they are placed within a patch, but replicates are placed in different patches (Figure 10b) there will be positive correlation among treatments.

This contrasts with a typical, independently sampled experiment as shown in Figure 10a where control and experimental plots are scattered over the study-site.

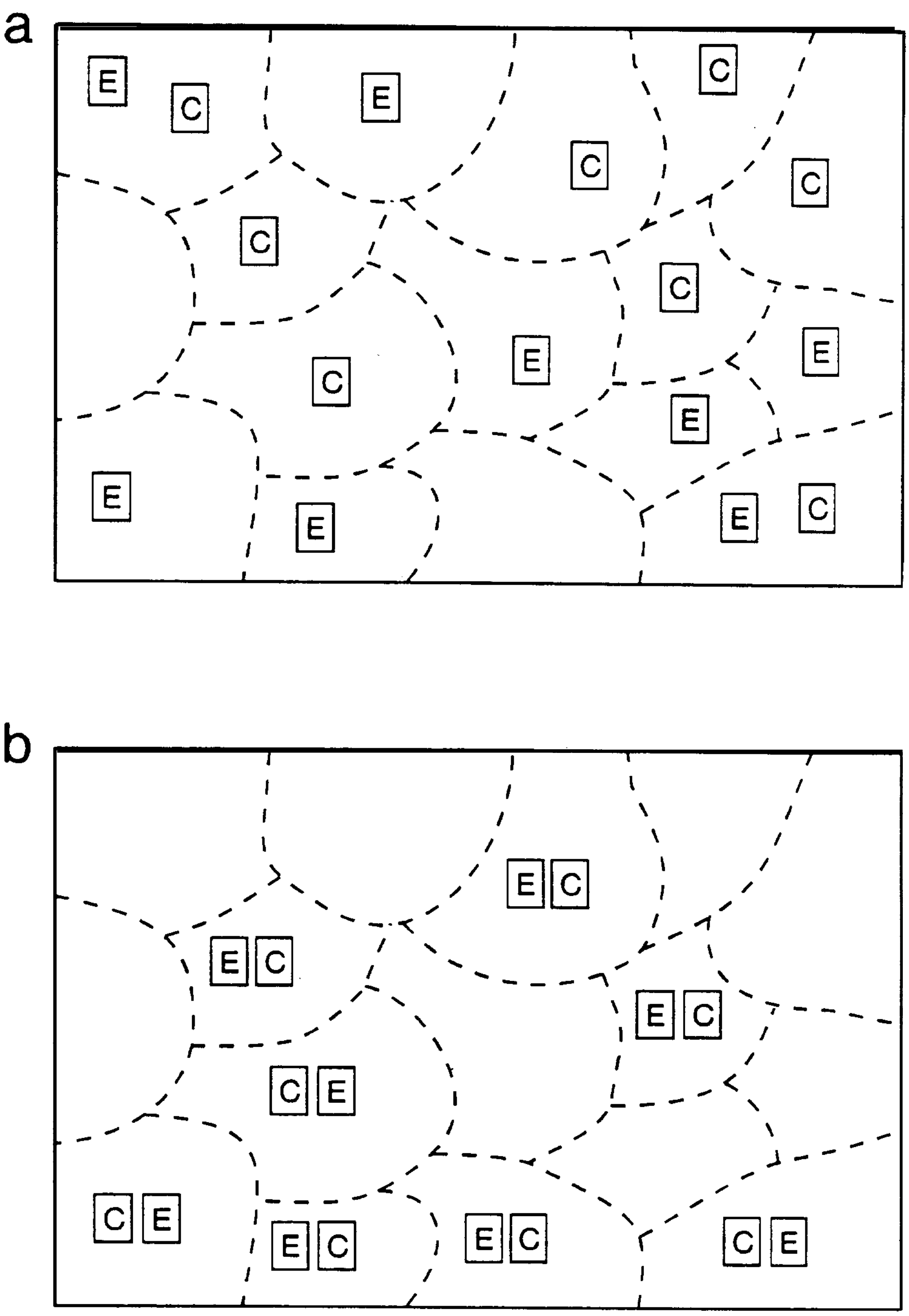

Figure 10. Comparison of control (untouched) and experimental (prey species 2 removed) plots to test an hypothesis about effects of predation on prey species 1. There are 8 replicates of each treatment. In (a), replicates are scattered randomly across the study-site, independently within and between treatments, In (b), a replicate of each treatment is placed close together (in the mistaken belief that this “reduces” variability). In both diagrams, the dotted lines indicate patches of habitat with different intrinsic rates of predation - the arrangement, boundaries and other details of which are not known to the experimenter. C, E represent control and experimental plots, respectively.

In the first experimental design, the variability among replicates consists of variations from patch to patch and the individual differences among plots in a patch. In contrast, for the control and experimental plots in the same patch, the only difference is the smaller, within-patch variation. The variation among treatments no longer includes any component measuring variability among patches! The effect of this is that variability among treatments, if the hypothesis is false and there is no average difference due to the experimental treatments, will be smaller than that among replicates within treatments. F-ratios will be smaller and the probability of Type I was much less than stated by a table of F-ratios. At the same time, the probability of detecting the difference (the power of the experiment) will markedly decline.

It is not always possible to do formal tests on experimental results to determine whether there are patterns of non-independence. The sort of issues raised by non-independence among treatments can best be dealt with by two considerations. First, a considerable amount more thought should be given to the biology and processes operating in the system being studied. Being aware of the need to avoid non-independence is the most important step. Don’t use the same plots, animals, tanks, nests, etc., in more than one treatment if they are the replicated units. At some scales, there is no prospect of eliminating correlations.

Within treatments, among replicates, it is sometimes possible to test for correlations prior to analysis. Where large numbers of repeated readings are taken on the behaviour or physiology or growth of individuals, these can be examined by temporal auto-correlations. The analysis can be used to determine at what temporal scales data are sufficiently independent to be used as independent replicates. The same procedures can be used where there are numerous spatial replicates, such as transects of continuously monitored plankton or oceanographic variables.

4.6.2 **Heterogeneity of Variances**

The second assumption - that variances within samples were equal - has consequences for the validity of the analysis when there are marked departures from homogeneity of the variances. The distribution of the *F*-statistic was calculated under the assumption that all populations have the same variance. When they do not, the probability of Type I error (rejecting the null hypothesis even though the populations have the same mean) is much greater than specified for any chosen critical value. This is only true when sampling is balanced (i.e. all the samples have the same number of replicates). When samples are not the same size, heterogeneity of variances can cause decreased probability of Type I error.

It turns out that, for balanced samples, this problem is serious for use of analysis of variance when one of the sampled populations has a larger variance than the others. It is much less of a problem, or no problem, when the variances differ from one population to another, but there is no single exceptionally large variance. It is not a problem when one variance is smaller than the others (e.g. Box 1953).

Also, the effects of heterogeneity of variances are much worse if the sizes of samples differ from one population to another. This suggests very strongly that samples should be balanced.

Several tests have been proposed and are widely used in the ecological literature. Cochran’s test is probably the most useful. It uses as a test statistic the ratio of the largest variance to the sum of the sampled variances.

$$\text{Cochran's} \quad C \quad \frac{32\text{largests;}}{\sum_{i=1}^{a} s_i^2}$$

The frequency distribution of *C* has been tabulated when the null hypothesis is true and the variances are equal. The table of *C* involves *a*, the number of treatments or populations and *(n* - 1), the degrees of freedom i n each sample. Note that *n* must be the same for all samples. If Cochran’s test is significant, there is evidence of a serious potential problem for any ensuing analysis of variance.

One of the first things to consider when experimental data have heterogeneous variances is “why”? The first step in dealing with the discovery of heterogeneous variances is to determine whether this is unusual, If you have experience with the type of data you are analysing and your previous experience suggests that [here is normally no heterogeneity of variances. You should give careful thought to the individual replicates in the sample that generated the excessive variance. The data may be wrong,

If there is evidence of an error, the erroneous datum should be omitted from the analysis and replaced with the average of the other replicates. This dummy number has the property that it keeps the set of data balanced, but does not alter the mean of the valid data, nor alter their variance. Analyse the data and adjust the degrees of freedom for the within samples sum of squares by subtracting one (for the missing or dummy replicate).

If a complete set of replicates is wrong, the sample could be eliminated from the analysis. Calculate the sum of squares for that sample (i.e. the variance of the sample multiplied by its degrees of freedom) and subtract it from the sum of squares within samples in the analysis of variance. Then subtract the degrees of freedom for that sample from the degrees of freedom within samples. Now you can calculate a new mean square within samples that does not use the large variance from the odd sample,

You cannot remove offending replicates or samples from your data just because they cause heterogeneity of variances. There must be justifiable cause based on the previous, extensive sets of data available from similar work to demonstrate that an error has been made.

4.63 Transformations of Data

An important procedure for dealing with heterogeneity of variances is to transform the data to some scale in which there is no heterogeneity. This is only effective if there is a relatively constant relationship between variances of samples and their means, so that where there are differences among means, there will be differences among the variances. Furthermore, to retain the possibility of being able to interpret the transformed data, transformations must be **monotonic.** This means that the transformation leaves the means of different samples in the same rank order (the largest is still the largest; any mean larger than another in untransformed scale will still be larger than the second one when transformed).

Many types of biological data, particularly those involving frequencies or counts per unit are distributed approximately as Poisson distributions. It is a property of these distributions that their variances equal their means. So, in any experiment where means differ, so will variances, thus compromising an analysis of variance.

In this case, the appropriate transformation is the square-root of the data. A preliminary step is to plot the variances against the means, which should be an approximately linear relationship. Transformation to $\sqrt{X}$ or $\sqrt{}(X + 1)$ generally removes the heterogeneity

Many types of data, particularly data that are ratios of two variables, are highly skewed to the right. In such distributions, where the mean is large, the variance is very large. Such data are often approximately log-normally distributed. This situation can be seen in a graph of the standard deviations (i.e. square root of the variance) against the means. Where such a plot is approximately linear, transforming all the data to logarithms will usually remove the heterogeneity of variances. It is customary to add a small amount (1 or 0.1) to the numbers before transformation, especially if any data a zero. because the logarithm of zero is minus infinity.

Where data are percentages or proportions. [hey are often binomially distributed. As a result, variances are larger where means are near 0.5 (or 50%) than where means are small or large (near 0.1 (10 %) or 0.9 (90 %)). This can lead to heterogeneity of variances where means are different. In this case, the appropriate transformation is one that ‘spreads out’ the data towards the end of the possible range of values (i.e. near 0 or 1 for proportions or 0 and 100 for percentages). In contrast, the transformation should do little to the values near 0.5 or 50%. The appropriate transform is the arc-sin of the square-root of the proportion (i.e. the angle which has its size equal to the square-root of the data), Thus, the transformed data are:

$$X' = \sin^{-1}\sqrt{X}$$

where *Xs* are the original data, as proportions (i.e. divided by 100 if percentages).

Often, biological data being what they are, there are no simple ways of dealing with heterogeneous variances. Neither the approach of checking variances against published values, nor prior experience, nor the use of monotonic transformation solves the problem.

Under these circumstances. you should note that, in large experiments, analyses of variance are robust to departures from the assumptions. In other words, the validity of the test and the probabilities associated with the F-ratio distribution are not affected much by violations of the assumption. This is particularly true where data are balanced (i.e. sizes of samples are all made the same) and where samples are relatively large. “Relatively large” is one of those delightfully vague terms, but having more than about five treatments, with n more than about six seems satisfactory from simulations.

Where there are many samples or treatments, there can only be a small effect of any one sample variance on the estimated variances provided by the mean square within treatments. So, large, balanced experiments, particularly with large samples in each treatment, will not cause problems for interpretation of an analysis with heterogeneous variances.

For small, heterogeneous experiments, the best advice is to do the analysis anyway. If it turns out that there are no significant differences among treatments, this is a valid conclusion for balanced samples. It is valid because heterogeneity of variances leads to increased probability of Type I error. Therefore, if no significant differences occur among the means, you cannot make a Type I error, which is an erroneous **rejection** of a null hypothesis.

The only problem occurs where there are small samples with heterogeneous variances and their means differ according to the analysis of variance. This may be an erroneous result and should only be accepted cautiously.

4.6.4 Normality of Data

The assumption that data are normally distributed is not very important even where there are quite small samples (Box 1953). The analysis of variance is quite robust to non-normality - in other words, its outcome and interpretation are not affected by the data being non-normal. Again, this is particularly the case where experiments are large (there are many treatments) and/or samples of each treatment are large. It is also the case where samples are balanced.

5. MULTIPLE COMPARISONS TO IDENTIFY THE ALTERNATIVE HYPOTHESIS

Once an analysis of variance has rejected the null hypothesis that a set of *a* means differ, we must identify which means differ, Where there are more than two treatments, the pattern of difference cannot usually be identified by inspection. There are too many possibilities, We need a procedure to identify which means are different so there should be an orderly procedure for comparing all the means, in a logical and ordered way, to define which alternative to the null hypothesis is most likely to apply. There are many such procedures and there is much potential confusion about how to choose among them,

The major difficulty comes about because any attempt to identify, unambiguously, a specific alternative to the null hypothesis must involve several comparisons among the means of the treatments. If each comparison is done with a pre-set probability of Type I error, α, the set of tests has a much greater probability of error (see Ryan 1959, Tukey 1949 and Hartley 1955). So, multiple comparisons are made with procedures that attempt to protect you against increased probability of Type I error (rejecting a null hypothesis when it is true).

What follows is an attempt to consider what to do under different circumstances. The problem, the issues and the different options have been discussed by numerous authors. Prominently among these are the studies of simulated sets of data by Einot and Gabriel (1975) and Ramsey (1978). A very fine summary, with yet more simulations, has been provided for ecologists by Day and Quinn (1989). They made specific recommendations for many types of circumstances. By and large, the following is consistent with their recommendations.

Most environmental sampling requires what are called **a posteriori** comparisons, which means that the analysis of variance has shown there are differences and now, after (or a posteriori) the analysis, the patterns of difference will be examined. Consider the hypothesis that mean abundance of animals differs seasonally, but we do not know which seasons may differ. To examine this, numbers of animals are sampled in all four seasons and an analysis of variance is done. If the analysis is significant, we now compare all the means to find out which seasons differed. A posterior comparisons are tests of all possible pairs of means, in a chosen order, so that, if it can be, the alternative to the null hypothesis will be defined. Suppose the six possible comparisons are made and the outcome is interpreted as: $\overline{X}_{\text{spring}} > \overline{X}_{\text{summer}}$, $\overline{X}_{\text{spring}} > \overline{X}_{\text{autumn}}$, $\overline{X}_{\text{spring}} >$

$\overline{X}_{\text{winter}}$, $\overline{X}_{\text{summer}} > \overline{X}_{\text{autumn}}$, $\overline{X}_{\text{summer}} > \overline{X}_{\text{winter}}$; $\overline{X}_{\text{autumn}}$, $\overline{X}_{\text{winter}}$ do not differ. This series of paired comparisons leads, uniquely, (o the following alternative to the null hypothesis:

$$\overline{X}_{\text{spring}} > \overline{X}_{\text{summer}} > \overline{X}_{\text{autumn}} = \overline{X}_{\text{winter}}$$

It is precisely this sort of alternative that *a* posterior multiple comparison procedures are designed to determine. There are many possible alternatives, but the procedures attempt to identify which actually applies.

An alternative to the null hypothesis can only be identified if the following conditions apply to the outcome of the multiple comparisons. Means of all treatments can be arranged in groups (including groups containing only one mean) such that there are:

(i) No differences among the means within a group;

(ii) Every mean in one group differs from all means in any other group

Otherwise, no clear-cut result has been obtained and no alternative to the null hypothesis has been identified. This point will be discussed later.

5.1 SNK PROCEDURE

To illustrate a sequential procedure, the Student-Newman-Keuls (SNK) test is used here. The example is about observations that densities of a particular species of crab vary across low, middle and upper parts of a mudflat and in pools and in patches of seagrass on the mudflat. The hypothesis being examined is that there are consistent differences in densities among the five habitats where the species is found. In each habitat, the density of crabs was sampled with $n = 6$ representative quadrats scattered across appropriate parts of the mudflat. The first hypothesis to be examined is that there are different abundances among the 5 habitats. The means and analysis are in Table 6.

The first step is to arrange the means in ascending order. This provides the sequence in which to test them in pairs. The logic is that the most likely to pair to be different are those most apart in ascending order. The next most likely differences must be in those means most different, except for the two most extreme ones (i.e. means 2 to 5 and 1 to 4 in the ranked, ascending order). Thus, the sequence for testing is 5 to 1, then 5 to 2, 4 to 1; then 5 to 3, 4 to 2 and 3 to 1. Finally, adjacent pairs are tested (5 to 4, 4 to 3, 3 to 2 and 2 to 1). This is the **sequence** of the test.

In each set of comparisons, what is needed is a standard error for the difference between two means. Here, because only balanced analyses (with *n* the same for all treatments) are considered, this is simply calculated. The standard error for each sample is the square-root of the sample variance divided by the size of sample (S.E. = $\sqrt{s^2/n}$). We have already concluded that the variances of the sampled populations are the same (it is one of the tested assumptions before we did the analysis of variance). So, we have a common estimate of variance, pooled from all samples, in the Mean Square within treatments. Thus:

$$\text{Standard Error} = \sqrt{\frac{\text{Mean Square Within Samples}}{\text{n}}}$$

with $a(n - 1)$ degrees of freedom, i.e. those associated with this mean square (Table 6).

Table 6

Numerical calculation of Student-Newman-Keuls' and Ryan's multiple comparisons. Data are densities of crabs in five habitats (n = 7 quadrats in each habitat).

a) Analysis of variance

Habitat	Lowshore	Midshore	Highshores	Pools	Patches of seagrass
Mean	6.4	7.1	3.8	2.6	4.1
s^2	2.18	0.67	1.77	1.35	2.06

Source of variation	Sum of squares	Degrees of freedom	Mean square	F- ratio	
Among Habitats	99.26	4	24.82	15.38	$P < 0.01$
Within Habitats	48.18	30	I .60		
Total	147.65	34			

Standard Error for means = $\sqrt{1.60 / 7} = 0.48$

b) Multiple comparisons. For SNK test, P is nominally 0.05. Any difference between two means larger than the relevant product Q x S.E. is significant at P = 0.05. Significance is denoted by *.

Rank order	1	2	3	4	5			
Ranked means	2.6	3.8	4.1	6.4	7.1	g	Q	Q x S.E.
Comparisons	5-1 4.5*					5	2.89	1.39
	4-1 3.8*	5-2 3.3*				4	3.49	1.68
	3-1 1.5	4-2 2.6*	5-3 3.0*			3	3.84	1.84
	2-1 [a]	3-2 [a]	4-3 2.3*	5-4 0.7		2	4.10	1.97

[a] Tests not done because 3 - 1 was not significant.

Conclusion: Mean density in lowshore and midshore areas is similar, but greater than in othcr habitats, which arc also different from each other.

For each pair to be compared, we need to calculate a test statistic,

$$Q_{i,j} = \frac{\overline{X}_i - \overline{X}_j}{\text{S.E.}}$$

where $\overline{X}_i$ and $\overline{X}_j$ are the two means being compared and S.E. is the standard error defined above. The distribution of Q is tabulated when the null hypothesis is true and there is no difference between two means(e.g. Snedecor and Cochran 1989; Winer et al. 1991) and its distribution depends on the range of number of means across which any two are compared and the degrees of freedom for the standard error (i.e. *a*(*n* - 1), see above). The range of number of means is five (g = 5 in Table 6) for the first comparison (smallest to largest means), four (g = 4 in Table 6) for the next step (second smallest to largest; smallest to second largest) and so on, as in Table 6.

So, the tests proceed sequentially from the furthest apart to the closest. At each step, the calculated values of Q are evaluated for significance. Where means do not differ, testing stops. So, in Table 6, because mean 3 is not significantly different from mean 1, the comparisons of means 1 and 2, and means 2 and 3 are not done. There is already evidence that a single group of means exists which includes the smallest three (1 and 3 do not differ). On the other hand, further tests are needed to determine whether the largest two means (4 and 5) differ, so the final row of tests is completed for this comparison.

5.2 INTERPRETATION OF THE TESTS

The procedure is straightforward; the rules are simple. In this case, there is a clear interpretation and an obvious alternative to the null hypothesis has been identified. This will not always happen. The multiple comparisons are not as powerful as the original F-test in the analysis of variance. So, there are going to be cases where the analysis of variance causes rejection of the null hypothesis, but the multiple comparisons do not allow identification of a valid alternative to it. This is not at all unreasonable. It is a far simpler matter, requiring far less information, to determine that a null hypothesis is false and therefore one of a set of many possible alternatives applies than it is to identify which particular alternative applies. Thus, sometimes the multiple comparisons do not show any differences.

The second failure to identify an alternative to the null hypothesis occurs when contradictions are caused by the multiple comparisons. Suppose that instead of a clear-cut pattern (as in the example in Table 6), you get the results in Table 7. There is no doubt that means 1 -3 are smaller than means 5 and 6. Nor is there doubt that means 1 -3 are similar and means 5 and 6 are similar. The problem is mean 4. It is different from the group containing 1, 2 and 3 because it differs from 1 and 2. It does not, however, differ from 3. At the same time, it cannot be included in a group with 6, but is not different from 5. The result is not interpretable (as in Table 7). The procedure has failed to provide an alternative to the null hypothesis.

Table 7

Illustration of Student-Newman-Keuls' multiple comparison producing an illogical result. Means are from 6 experimental groups, with $n = 6$ replicates

a) Data and analysis of variance

Treatment	1	2	3	4	5	6
Mean	59.8	52.7	66.7	73.4	48.2	50.3

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F*- ratio	
Among Treatments	2987.45	5	597.49	29.42	$P < 0.001$
Within Treatments	609.30	30	20.3 I		
Total	3596.75	35			

Standard Error for means = $\sqrt{20.31/6} = 1.84$

b) SNK test. Apparently significant differences at $P = 0.05$ arc shown as *.

Rank order	1	2	3	4	5	6			
Ranked means	48.2	50.3	52.7	59.8	66.7	73.4	*g*	*Q*	*Q* x S.E.
Comparisons	$^{6-1}$ 25.2*						6	2.89	5.32
	$^{5-1}$ 18.5*	$^{6-2}$ 23.1*					5	3.49	6.42
	$^{4-1}$ 11.6*	$^{5-2}$ 16.4*	$^{6-3}$ 20.7*				4	3.84	7.07
	$^{3-1}$ 4.5	$^{4-2}$ 9.5*	$^{5-3}$ 14.0*	$^{6-4}$ 13.6*			3	4.10	7.54
		$^{4-3}$ 7.1	$^{5-4}$ 6.9	$^{6-5}$ 6.7			2	4.30	7.91

Conclusion : Although treatment 4 (rank 6) has a larger mean than treatment 5 (rank 1), it is impossible to determine consistent groupings. Horizontal lines underline treatments that do not differ:

Rank	1	2	3	4	5	6
Treatment	5	6	2	1	3	4

Although commonly described, this *cannot* present a logical alternative to the null hypothesis!

What can be concluded is that there are differences, but the precise alternative hypothesis has not been identified, More experimentation is needed with increased sizes of samples because increasing *n* decreases the standard errors and decreases the sizes of critical values of *Q* because degrees of freedom are $a(n -1)$ for the standard error. Or, the results provide clues as to what may be happening, so that more refined or precise hypotheses can be proposed for subsequent experiments, The pattern of rank order of the mean will provide the new observations for new hypotheses and experiments, In particular, these might now have a priori determined statements about the alternative hypothesis and therefore more powerful multiple comparisons to detect it.

As a final point. multiple comparisons are often done when they are irrelevant. If, for example, the experiment is to test an hypothesis that rate of predation on mites differs at different distances from the edge of a fleld, then rate is measured in replicate plots at specified distances. The relationship should then best be analysed as a regression of rate of predation (Y) against distance (X). There is no need for means at adjacent distances to be different (as is the pattern sought by multiple comparisons), The relationship between predation and distance is what needs to be examined. It is not at all uncommon for multiple comparisons to be used under such inappropriate circumstances (see Dawkins, 1981). Adjacent distances or any adjacent values of *X* in any regression do not have to differ, even if the relationship of some variable with X is significant, The logical basis of multiple comparisons must not be lost in the rush to use them.

6. FIXED OR RANDOM FACTORS

One very important consideration in an analysis of variance is to be sure whether experimental treatments represent **freed** or **random** factors. This makes little practical difference for a one-factor analysis, but will turn out to matter a great deal for extensions beyond one factor.

The relevant definition of an experimental factor is a function of the way hypotheses are stated and **not** a property of the analysis. As an immediate consequence of this, you should always know which sort of factor you have because you cannot interpret the experiment without the hypothesis being clear.

The difference revolves around whether or not you are concerned with specific, identifiable required treatments (which would make a fixed factor) or whether you are interested in a general problem, some components of which are included (representatively) in the experiment. This second choice would make it a random factor.

Here is a specific example to illustrate the nature of the difference. In studies of the fecundity of several species of fish, there is variation from fish to fish, but it has been proposed that there are also differences among species. This hypothesis can be tested by measuring fecundity in a number of fish of several species. The null hypothesis is that there will be no difference in mean fecundity among species. The species chosen are the experimental factor and the individual fish are the replicates.

There are two ways you might decide which species to choose, These “ways” encapsulate the different sorts of hypothesis. In one case, there is no reasoning behind which species to choose. The hypothesis is very general (“species differ”) and you are examining four species in order to get a range of species. You have no particular reason for choosing any particular species. You choose four essentially at random out of those available. In [his case, the species are randomly chosen to represent [he population of the species in the area. Any set of four species will do the job - the hypothesis simply considers that there are differences among species,

In this case, the hypothesis simply says "there are variations among species" and therefore the difference between [he null hypothesis ("no difference among species") and the hypothesis is whether or not there is significant variation. If there is, the null hypothesis will be rejected - but you would also expect to reject it for any set of four species. In other words, the experimental factor consists of four representative species and any other sample of four species should be similarly representative. If one set of four shows variation, so should another set.

Contrast this with the following scenario. There are numerous species, but you only choose certain ones to examine. You look at A and B because they are the ones most discussed in the literature. You choose Species C because it is the most abundant in the area. Finally, you choose Species D because you are concerned about its Iong-term conservation and so you would like to include it in your study,

Now you no longer have a random factor. You have, in the way you are now phrasing the hypothesis, created a fixed factor. The levels of the factor, i.e. the particular species chosen, are fixed by special characteristics they have (or that you use to describe them). You are no longer representing all species - you are specifying a particular set of species and they cannot. represent any others. The species chosen are the complete set of species specified - they are not a sample of a more generally defined population of species.

The example indicates that there are two different types of factor. In one - a fixed factor - all the relevant levels (i.e. relevant to the hypothesis being tested) are included in the experiment. The experiment provides a precise rationale for rejecting or retaining the null hypothesis as it applies to those treatments.

The other type of factor - a random factor - includes only a sample of relevant levels. The population of levels or treatments to which they hypothesis applies is much larger than the number of levels included. The experiment provides a less precise answer. Its results apply to a sample; they are, in a sense, a sampled answer. As with any sample, the answer represents the answer for the population and is not exactly the answer for all possible levels that might have been examined. On the other hand, this loss of precision is compensated by having a more general answer. The outcome of the experiment should, all other things being equal, be the outcome that would be obtained in experiments on other samples of levels that might have been included.

There are several other ways of considering this concept. A useful rule-of-thumb is to decide, if you were going to repeat the experiment as a new test of the hypothesis, would YOU have TO use the same levels of the treatment (Simpson et al. 1960)? If any set of treatments (e.g., in this case, species) will do, they are a random factor. For the fixed factor situation you **must** use the same set of treatments, otherwise you would clearly be testing a different hypothesis.

Again, consider a test of the hypothesis that abundance of a population changes from time to time. For example, under the model that disturbances lead to fluctuations in numbers, this would be a relevant hypothesis provided that the interval between times is long enough for disturbances to occur. .At each time, several replicate samples are taken to estimate mean abundance. These are compared by a one-factor analysis of variance with Times as the factor. The times chosen can be entirely random. All that matters is to determine whether abundance varies. Someone else testing this hypothesis can use completely different times of sampling and should expect to get a similar answer,

Contrast this with the situation where change in abundance of a population is explained as a result of particular processes operating after a disturbance, so that there must be an increase in numbers between the second and third year after a disturbance. Now, the hypothesis predicts a particular pattern in abundance *and* when it should occur. The time elapsed since the disturbance is important and samples must be taken at the correct times. This fixes the times and the hypothesis also fixes how many times must be sampled. In an analysis, Time is now a fixed factor. Anyone else doing such an experiment must use the same times of sampling after a disturbance. The only times that matter for the hypothesis must be sampled in the experiment.

6.1 INTERPRETATION OF FIXED OR RANDOM FACTORS

Obviously, the definition and interpretation of experimental factors depend entirely on the null hypothesis and hypothesis being examined. There is, however, a major difference between the interpretations that are appropriate for random as opposed to fixed factors, Consider the fixed case first. The whole notion is that there is a series of levels of a factor chosen because they are needed by the null hypothesis. All of the treatments are equally necessary and the alternative to the null hypothesis must involve a consideration of all the levels of the factor.

In contrast, a random factor is **sampled** in the experiment. A number of levels (places, sites, species, whatever) is chosen to represent those available and defined as relevant by the null hypothesis, Differences among them, so that the null hypothesis is rejected, indicate that there are, in general, differences among samples of the levels of the factor. The individual treatments do not matter. Comparisons among the specified levels used in the experiment are not necessary, nor, usually, useful. Discovering that Place A differs from B, C and D and that C differs from D cannot be particularly informative. After all, to test the null hypothesis of no differences among places, the particular places A, B, C, and D might not even have been examined because they might not have been sampled as randomly chosen levels of a factor. Thus, the information that there **are** differences is all that is needed. The specific pattern of differences does not help in any interpretation - it was not part of an hypothesis.

7. NESTED OR HIERARCHICAL DESIGNS

I\ Tested or hierarchical designs of sampling and experiments are used in environmental work to ensure appropriate replication. Often, it is impossible to sample all of a series of replicated locations, *so* representative areas are chosen and each is sampled by a series of quadrats. This gives data for a set of locations, areas in each location and quadrats in each region - a clear hierarchy. We say that the quadrats are nested in each area - they can only be in one area. The areas are nested in a location - each area can only be in one location. The main use of such sampling in environmental work is to collect data from several spatial scales or several times scales (see later).

A nested experimental design is one that uses replication of experimental units in at least two levels of a hierarchy. In the above example, these are locations, areas and quadrats. These designs are widespread in environmental work and are extensions of the one factor experiment considered earlier,

Hurlbert (1984) reviewed a range of different types of ecological experiments to discuss their shortcomings in terms of inadequate (or no) replication. Anyone interested in recognizing mistakes in the design and interpretation of experiments should read his paper.

7.1 THE ANALYSIS OF VARIANCE

Data collected in a nested or hierarchical design are replicated samples (each of size *n)* from the replicated plots or aquaria or whatever sample units (with *b* replicated plots in each experimental treatment). In the case of the spatial sampling earlier, there would be *n* quadrats in each of *b* areas in each of a locations.

The data would be as in Table 8. This has a total amount of variability, measured exactly as in the one-factor case by the sum of squared deviations of all data from the overall mean of the entire set. The only difference from the earlier case is that variability must be added up over all *n* replicates, all *b* plots and all *a* treatments, i.e. over three levels of variability.

$$\text{Total Sum of Squares} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(X_{ijk}-\bar{X}\right)^2$$

The total variability can be partitioned into three components:

$$\text{Total Sum of Squares} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(X_{ijk}-\bar{X}_{j(i)}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{X}_{j(i)}-\bar{X}_i\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{X}_i-\bar{X}\right)^2$$

(i) (ii) (iii)

These three terms are, without much difficulty, identifiable in terms of the data, as:

(i) variation among replicates (X_{ijk}) in an area, as deviations from the mean (~,(,)) for a given area *(j)* in a particular location (i). This is variability within areas and is exactly equivalent to the measure of variation within samples in the one-factor case, except there are now *b* samples (one for each unit) in each of the a treatments.

(ii) variation among areas in each location. The means of each of the *b* areas in a given location (i) are $\bar{X}_{j(i)}$ · This second component of variation measures how different the areas are from the average number in location i ($\bar{X}_i$). If areas do not differ from one another, the mean of any variable measured in them should be the same, except for sampling error. So, in each treatment, this term measures the variability among areas within locations.

(iii) variation among locations, exactly as in the one-factor case, except that the mean value in each treatment is estimated from *n* replicates in each of *b areas,* rather than from a single sample.

Table 8

Data for an analysis of the numbers of polychaete worms in three locations *(a* = 3), each sampled in *b* = 5 areas. In each area, *n* = 6 quadrats are sampled and the number of worms recorded. Data are X_{ijk} where *i* is the location, *j(i)* is the area in location *i* and *k is* the quadrat in area *j* in location *i.*

Location *(i)*				1					2					3		
Area *(j)*		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Quadrat *(k)*																
	1	82	79	90	75	38	92	62	67	95	70	74	47	60	43	47
	2	67	84	100	93	64	80	97	64	93	62	76	71	88	53	44
	3	73	70	65	99	80	83	63	85	100	77	72	54	86	48	16
	4	70	71	99	95	74	71	77	83	80	80	71	56	84	79	43
	5	83	67	84	92	87	52	88	79	83	71	60	77	45	70	49
	6	95	80	63	95	79	73	77	88	76	87	74	66	48	45	55
Mean ($\overline{X}_{j(i)}$)		78.3	75.2	83.5	91.5	70.3	76.2	77.3	77.7	87.8	74.5	71.2	61.8	68.5	56.3	42.3
Variance		108.0	45.4	264.0	71.1	309.0	181.0	188.0	98.3	90.2	76.3	33.0	129.0	394.0	217.0	185.0
Mean ($\overline{X}_i$)				79.8					78.7					60.0		
Mean ($\overline{X}$)									72.8							

Table 9

Nested analysis of variance of an experiment with a locations (1...*i*...*a*), each with *b* areas (1.. .*j*...*b* in each *i)* and each sampled with *n* quadrats (1...*k*...*n* in each *j(i)).*

Source of variation	Sums of squares	Degrees of freedom	Mean square	Mean square estimates
Among Locations = A	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{X}_i - \bar{X}\right)^2 = SS_A$	$((a - 1)$	$MS_A = SS_A / (a - 1)$	$\sigma_e^2 + n\sigma_B^2 + bn\sigma_e^2$
Among Areas in each Location = B(A)	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{X}_{j(i)} - \bar{X}_i\right)^2 = SS_{B(A)}$	$a(b - 1)$	$MS_{B(A)} = SS_{B(A)} / a(b - 1)$	$\sigma_e^2 + n\sigma_B^2$
Within areas	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{X}_{ijk} - \bar{X}_{j(i)}\right)^2 = SS_W$	$ab(n - 1)$	$MS_W = SS_W / ab(n - 1)$	σ_e^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{X}_{ijk} - \bar{X}\right)^2$	$abn - 1$		

The data in the example in Table 8 are analysed using these formulae (Tables 9 and 10). The actual calculations are according to the “machine formulae” - simpler, but algebraically identical forms which do not introduce rounding errors.

Table 10

Nested analysis of variance of number of worms (in Table 8),

Cochran’s test for heterogeneity of variances, $C = \frac{394}{2311} = 0.17$, $P > 0.05$

Analysis of variance:

Source of variation	Sum of Squares	Degrees of freedom	Mean square	F-ratio	
Among Locations	6865.39	2	3432.70	8.32	$P < 0.01$
Among Areas(Locations)	4951.39	12	412.62		
Within areas	11552.64	75	15404		
Total	23369.43	89			

7.2 THE LINEAR MODEL

As with the single-factor case, a linear model can be used to describe the data.

$$X_{ijk} = \mu + A_i + B_{j(i)} + e_{ijk}$$

where X_{ijk} is any replicate (k) in any area *(j)* in a given location (i).

The $B_{j(i)}$ terms represent the differences in means of populations in each of the *j* replicate areas nested in the i experimental locations. Thus, if the number of worms differ from area to area, the $B_{j(i)}$ terms indicate these differences. In contrast, if there are no differences among the means in the areas, the $B_{j(i)}$ terms are zero. The final, e_{ijk} term is exactly as before. It represents the individual variability in numbers from one quadrat *(k)* to another in a particular area. The experimental design is hierarchical because it has a hierarchy of three spatial scales.

The analysis of variance of the nested model is completed, as for the previous case, by determining the degrees of freedom associated with each sum of squares and then calculating the Mean Squares. These are shown in Table 9 and there is a worked example in Table 10, using the data in Table 8.

The analysis allows estimates of the variation among quadrats in each area (σ_e^2 in Table 9). This is as before for the one-factor case. Also estimated is $\sigma^2_{B(A)}$ which measures variation among areas in each location. If the distribution of worms if fairly even over a location, there will be no differences from area to area and the variance among areas will be zero.

Finally, the analysis estimates variation among locations (σ^2_A). If locations differ in mean number of worms, this variance measures the differences. If the locations do not differ, this variance will be zero. So, the null hypothesis that locations have the same mean number of worms:

$$H_0: \quad \mu_1 = \mu_2 \ldots\ldots = \mu_i \ldots\ldots = \mu_a \quad (= \mu)$$

is the same as :

$$H_0 : \sigma^2_A = 0$$

If there are differences among locations, the Mean Squares for differences among locations (MS_A) will be larger than the Mean Square for differences among areas ($MS_{B(A)}$). If instead the null hypothesis is true, these should be the same, except for sampling error. Thus:

$$F = \text{Mean Square}_A \,/\, \text{Mean Square}_{B(A)}$$

with *(a* - 1) and *a(b* - 1) degrees of freedom can be used to test the null hypothesis. If *F is* larger than 1 it implies that locations differ. Note that this result is true whether or' not there is variation among areas in each location. So, if areas do not differ, $\sigma^2_{B(A)}$ will be zero, but the above test is still valid for testing for differences among locations.

7.3 MULTIPLE COMPARISONS

When the null hypothesis is rejected, multiple comparisons are needed on the means of the experimental treatments. These are exactly as described for the single-factor case, except that care must be taken to ensure that the correct standard errors and degrees of freedom are used. Instead of the Within Sample mean square being used to construct standard errors, the appropriate mean square is that among units within each treatment (i.e. B(A) in Table 9). This is used to construct standard errors because it includes all sources of variation within each treatment.

An example using Student-Newman-Keuls tests is provided in Table 11. Nothing further needs to be discussed here.

7.4 COST BENEFIT OPTIMIZATION OF SAMPLING DESIGNS

One useful property of nested sampling designs is that they provide exact solutions for optimizing the use of replication at the various levels included in the experiment. Consider the simplest case of sampling several locations in an estuary to detect differences from place to place in the concentration of a toxic chemical. Sampling will be done in a = 3 locations about 1 km apart and about 200 m square.. In each, b = 2 areas are sampled, about 50 m apart and 30 m square. In each *area, n* = 3 cores are taken to analyse the chemicals.

Optimization of replication is appropriate to ensure that the number of replicates per area (n) and the number of areas per location *(b)* are chosen to maximize the precision of the estimates for the minimal possible cost in terms of money per replicate or per area. The principles of this cost-benefit procedure are straightforward. The ideal is to maximize the precision with which means are estimated for the minimal possible number of samples (i.e. cost of experiment). Thus, the best designs have the smallest standard errors for the smallest possible cost. Examples of calculations for real experiments in marine ecology are available in Kennelly and Underwood (1984, 1985).

Table 11

Student-Newman-Keuls multiple comparison of means after a significant nested analysis of variance. Data are analysed in Table 10. The standard error for the means of the three locations is calculated from the mean square among areas within each treatment. Q values are for probability of Type I error = 0.05, with g and 12 degrees of freedom (i.e. degrees of freedom for the mean square used to calculate the standard error).

Location	1	2	3
Mean	79.8	78,7	60.0

$$\text{Standard error} = \sqrt{\frac{\text{Mean square Plots (Treatments)}}{\text{No. of data in each mean}}} = \sqrt{\frac{412.62}{30}} = 3.71$$

Rank Order	1	2	3	*g*	*Q*	*Q* x S.E
	60.0	78.7	79.8			
	3-1 19.8 *			3	3.41	12.65
	2-1 18.7 *	3-2 1.1		2	2.84	10.54

Conclusion: Location 1 had fewer worms per unit area.

In the case of a nested design, there are two sources of variation in the standard errors (σ_e^2 and $\sigma^2_{B(A)}$ both contribute) and two costs of replication (the cost per replicate X_{ijk} and the cost of sampling the **b** replicated units). The variance associated with estimated means of locations is:

Variance . (Standard Error)*

$$= \frac{\text{Variation among Areas}}{\text{Size of sample used to estimate means of areas}}$$

$$\frac{\text{Mean Square Among Areas}}{\text{No. areas X No. replicates in each area}}$$

The Mean Square among areas (B(A) in Table 12) is the variance within the treatments it includes the variance among replicates in each unit and the variance among units. The mean of each treatment is averaged over *n* replicates in each of *b* units. Therefore the variance of estimated means of locations is:

$$\text{Variance} = V \text{ estimates } \frac{\sigma_e^2 + n\sigma^2_{B(A)}}{bn}$$

The cost of the experiment can be estimated or measured in accordance with the following:

Cost of each treatment = C = *brick* + bc_j

where c_k is the cost of each of the *n* replicates (i.e. of each of the *bn* cores in all *b* areas in a location) and c_j is *the* cost of sampling each of the *b* areas. To design the most efficient experiment, the product of these must be kept as small as possible, so:

$$VC = c_k\sigma_e^2 + \frac{c_j\sigma_e^2}{n} + nc_k\sigma^2_{B(A)} + c_j\sigma^2_{B(A)}$$

The minimal value of this product can be found by first differentiating with respect to *n* and setting the differential to zero to identify its minimum, which gives:

$$n = \sqrt{\frac{c_j\sigma_e^2}{c_k\sigma^2_{B(A)}}}$$

If we obtain estimates of the costs and the variances, we can calculate the optimal number of replicates per area. (n). The costs are obtained from the literature, previous similar experimental and quantitative work and, as discussed below, pilot studies. The estimates of variance can sometimes be found in the literature, from previous experience and estimates or from pilot studies. Assume some preliminary work has already been done to allow the calculation of mean squares as in a nested analysis, Specifically, assume a small experiment has been done with *n* = 3 replicate cores in *b* = 2 replicate areas for these 3 locations. The analysis of variance is in Table 12, where estimates of costs and variance are calculated,

Once the above formula has been used to determine *n,* it is necessary to calculate *b.* There are two options for doing this, depending on whether the crucial limiting factor in the design of the experiment is its total cost or the precision which must be achieved.

Of the two, being limited by the total time or money (or other limiting cost) is more usual in environmental sampling and is easier to understand, The total resources available for the experiment must be divided among the *a* locations. The total resources allocated or available to the experiment divided by the number of treatments gives *C* in the Equation describing costs. The only unknown in the equation (once the costs per replicate and per area have been ascertained) is *b,* because *n* comes from the Cost-Benefit equation described above. Solving for *b is* straightforward as in Table 12. The cost-benefit procedure thus provides the optimal allocation of effort to *b* and to *n,* according to their relative variances and costs. In this case, there should be seven replicates in each of three replicated units in each treatment (i.e a total of 9 units and 63 replicate plants). The cost of the experiment is now *C* = $555 per treatment and a total cost of $1665.

Sometimes, of course, the optimization of n and *b* according to a set standard error will result in fewer experimental units being needed than was originally guessed - so that an experiment would be cheaper than found by arbitrarily allocating resources to it and solving for *b* using the cost equation, All this suggests that planning the sampling with some quantitative goals (in the form of precision needed or power to detect specified alternative hypotheses) will always be a more realistic way of finding out about nature.

Table 12

Example calculation of costs and variances per treatment in a nested analysis of variance with a = 3 locations , b = 2 replicated areas per location and n = 3 replicate quadrats per area.

a) Analysis of variance of data from pilot study.

Source of variation	Df	Mean Square	MS estimates
Among Locations = A	2	138.3	
Among Areas(Locations) = B(A)	3	103.1	$\sigma_e^2 + 3\sigma_{B(A)}^2$
Within areas	12	48.7	σ_e^2
Total	17		-

Costs: per replicate $=C_k=\$11.00$; per unit = $C_j= =\$108$

b) Determining n

Variance within areas = σ_e^2 estimated as 48,7

Variance among areas =

$\sigma_{B(A)}^2$ estimated as $\dfrac{\text{Mean Square B(A) - Mean Square Within}}{n}$ (from MS estimates)

= (103.1 - 48.7)/3 = 18.1

$$\therefore \quad n = \sqrt{\frac{(108)(48.7)}{(11)(18.1)}} = 6.3 \approx 7$$

c) Determining b

Using cost equation: C = \$600 available per location= $bnC_k + bC_j$

$= (b \times 7 \times 11) + (b \times 108)$

$\therefore b = 600/\,185 = 3.3 \approx 3$ to keep it within the available cost.

8. FACTORLAL DESIGNS

Factorial studies are investigations of more than one experimental treatment examined simultaneously. Examples are sampling programmes to deter-mine seasonal patterns of concentration of an enzyme in the blood of fish in several areas of coastline.

There is one fixed factor - the different seasons of the year with four levels. There is also a second factor, the places being sampled. This has a number of levels chosen by the experimenter and may be a fixed factor, comparing particular places, or a random factor comparing a representative sample of places to determine how much spatial variation there is. This is a factorial experiment. It includes a combination of two different experimental factors.

In general, factorial experiments are designed to be **orthogonal.** Orthogonality is the property that every level of one factor is present in the experiment in combination with every level of the other factor (or combinations of levels of other factors if there are more than two others). Ensuring orthogonality is important in order to analyse interactions, as discussed below.

The analysis proceeds in a way very similar to those discussed earlier. There is a source of variation to compare the means of the first factor (A), with means $\overline{X}_1$, $\overline{X}_2 .... \overline{X}_i \overline{X}_a$ The sums of squares are

constructed exactly as in the one factor case. There is a corresponding term for means of levels of factor B (xl , $\overline{X}_2 \overline{X}_j X_b$), averaged over all levels of the first factor. So, one factor uses the average in each

season averaged over all of the places sampled. The second factor uses the average in all places averaged over all the seasons

The Residual term (Table 13) measures variation among replicates in each place and time. There is, however. one last term - the interaction between Places and Times (or, generally, A and B),

Table 13

Analysis of variance of an experiment with two fixed experimental factors, Factor *A* has levels 1...*i*...*a*; Factor *B* has levels 1...*j*...*b*. There are *n* replicates in each combination of levels *i* of Factor *A* and *j* of Factor *B*. Every set of replicates samples a population with variance σ_e^2 *(i.e.* all *ab* populations have homogeneous variances). k² terns measure differences among the levels of the various factors.

Source of variation	Sum of squares	Df	Mean square estimates
Among *A*	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{X}_i - \overline{X})^2$	a - 1	$\sigma_e^2 + bnk_A^2$
Among *B*	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{X}_j - \overline{X})^2$	b - 1	$\sigma_e^2 + ank_B^2$
$A \times B$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{X}_{ij} - \overline{X}_i - \overline{X}_j + \overline{X})^2$	$(a$ - 1$)(b$ - 1)	$\sigma_e^2 + nk_{AB}^2$
Residual	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \overline{X}_{ij})^2$	$ab(n$ - 1)	σ_e^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \overline{X})^2$	abn - 1	-

Any orthogonal combination of experimental treatments (however many factors are involved) will potentially cause an interaction in the analysis. The degrees of freedom for each interaction are the products of the degrees of freedom in the factors making the interaction.

In the case of two factors, there is only one possibie interaction - *A* x *B*. The degrees of freedom calculated from the formula used to calculate the sum of squares of the interaction are also useful to confirm that the partitioning done is appropriate and complete.

8.1 APPROPRIATE NULL HYPOTHESES FOR A TWO-FACTOR EXPERIMENT

The preceding account of how to partition the variability among the data allows a formal statement of relevant null hypotheses for the experiment. The partitioning was done on the basis that differences among levels of Factor *A* could be examined by ignoring any effects of Factor *B*. This requires that any influence of different levels of Factor B must be equal for all levels of Factor A. An alternative way of stating this is that the influence of Factor B (i.e. whether levels of B differ or not) must be independent of the differences among levels of A (if there are any). SO, for Factor A there are *a* populations with means $\mu_1, \mu_2 \mu_i \mu_a$ and:

$$H_{01}: \ \mu_1 = \mu_2 \ ..... \ = \mu_i \ ..... \ = \mu_a$$

assuming that there is independence between Factors *A* and *B*. Simultaneously, for Factor B:

$$H_{02}: \ \mu_1 = \mu_2 = \mu_j = \mu_b$$

assuming that there is independence between Factor B and Factor A.

The validity of posing and testing these null hypotheses obviously depends on the assumption of independence being true. To validate this assumption requires a test of the further hypothesis that Factors *A* and *B* are independent:

H_{03}: Differences among levels of Factor *A* are independent of differences among levels of Factor *B*

Any dependence between Factors A’ and B can be estimated from the mean square for the A x B interaction. It is therefore essential to test the interaction first. Then, if there is no evidence for dependence between A and B, the differences among levels of *A* and among levels of B are logically analyzable, Otherwise, they are not.

This is a matter of great importance because there is no logical validity in any test for differences among levels of either Factor if there is significant interaction between them. Note that construction of tests and interpretation of any factor in a multi-factorial analysis is totally and unambiguously dependent on there being no interaction between that source of variation and any other component of the analysis,

The exact form of the interaction term is complicated by whether either or both of Factors *A* and *B* are fixed or random. The analysis changes if one or other or both factors are random, as opposed to fixed, You can see the changes for possible sampling programmes in Table 14. The details of how to deal with this for any possible pattern are available in a number of text-books (Crowder and Hand 1990, Underwood 1981, Wirier et al. 1991). For many types of environmental sampling projects, the models are explained in full later in this guide-line, so you do not need to know the general principles - but it is a good idea to explore them if you want to understand more complex designs.

Table 9

Nested analysis of variance of an experiment with *a* locations (1...*i*...*a*), each with *b* areas (1..*j*...*b* in each *i*) and each sampled with *n* quadrats (1...*k*...*n* in each *j*(*i*)).

Source of variation	Sums of squares	Degrees of freedom	Mean square	Mean square estimates
Among Locations = A	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{X}_i-\overline{X})^2=\mathrm{SS_A}$	$(a - 1)$	$\mathrm{MS_A}=\mathrm{SS_A}/(a-1)$	$\sigma_e^2+n\sigma_B^2+bn\sigma_e^2$
Among Areas in each Location = B(A)	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{X}_{j(i)}-\overline{X}_i)^2=\mathrm{SS_{B(A)}}$	$a(b - 1)$	$\mathrm{MS_{B(A)}}=\mathrm{SS_{B(A)}}/a(b-1)$	$\sigma_e^2+n\sigma_B^2$
Within areas	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{X}_{ijk}-\overline{X}_{j(i)})^2=\mathrm{SS_W}$	$ab(n - 1)$	$\mathrm{MS_W}=\mathrm{SS_W}/ab(n-1)$	σ_e^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{X}_{ijk}-\overline{X})^2$	$abn - 1$	-	-

The main point to understand is that interactions exist wherever there are two or more orthogonal factors and the presence of interactions must be examined before testing for the main factors. Later, it will be explained that environmental impacts are interactions and interactions are, therefore, the focus for most statistical procedures to detect impacts.

8.2 MEANING AND INTERPRETATION OF INTERACTIONS

Understanding, interpreting and dealing with interactions among experimental factors is crucial to any sensible study of quantitative sampling. Consider an interaction between two fixed factors, *A* and *B* which have *a* and *b* levels, respectively. As an example, consider the investigation of concentrations of cadmium in tissues of animals in two areas in different seasons.

The sampling will therefore consist of a = 3 places (a fixed factor). There will also be b = 2 times of sampling, once before and once after an oil-spill in one of the places (P1). *n* replicate animals are sampled in each place and time. Figure 11 (a and b) illustrates the situation where there is no interaction. The amount of cadmium increased slightly between times 1 and 2 in all three places (Figure 1 lb). There was always more in Place 1 than the others (Figure 1lb), but no substantial increase there after the oil spill. You would conclude that there was more cadmium in one place than the others, a slight increase everywhere during the period of the oil spill, but no obvious pattern associated with the spill.

There is no interaction - the differences between the places are the same each time (Figure 1 lb), the difference from one time to the next are the same for each place (Figure 1la).

In contrast, consider the situation where there is an interaction. In Figure 11c, there is a large increase in cadmium from before to after the oil spill. This did not occur in the other two places. Looking at the data the other way, there is a large difference between P_1 and the other two places after the spill (at Time 2 in Figure 1ld) which was not there before. The interaction is obvious. Its interpretation as an environmental impact is clear. The increase happened in the appropriate place at the right time.

8.3 MULTIPLE COMPARISONS FOR TWO FACTORS

8.3.1 When There Is a Significant Interaction

If the interaction is significant, interest focuses on comparing the means of one factor separately at each level of the other factor and vice versa. The analyses available are then exactly the same as described for the single factorial experiment. If there have been *a priori* defined patterns of differences, *a priori* procedures should be used.

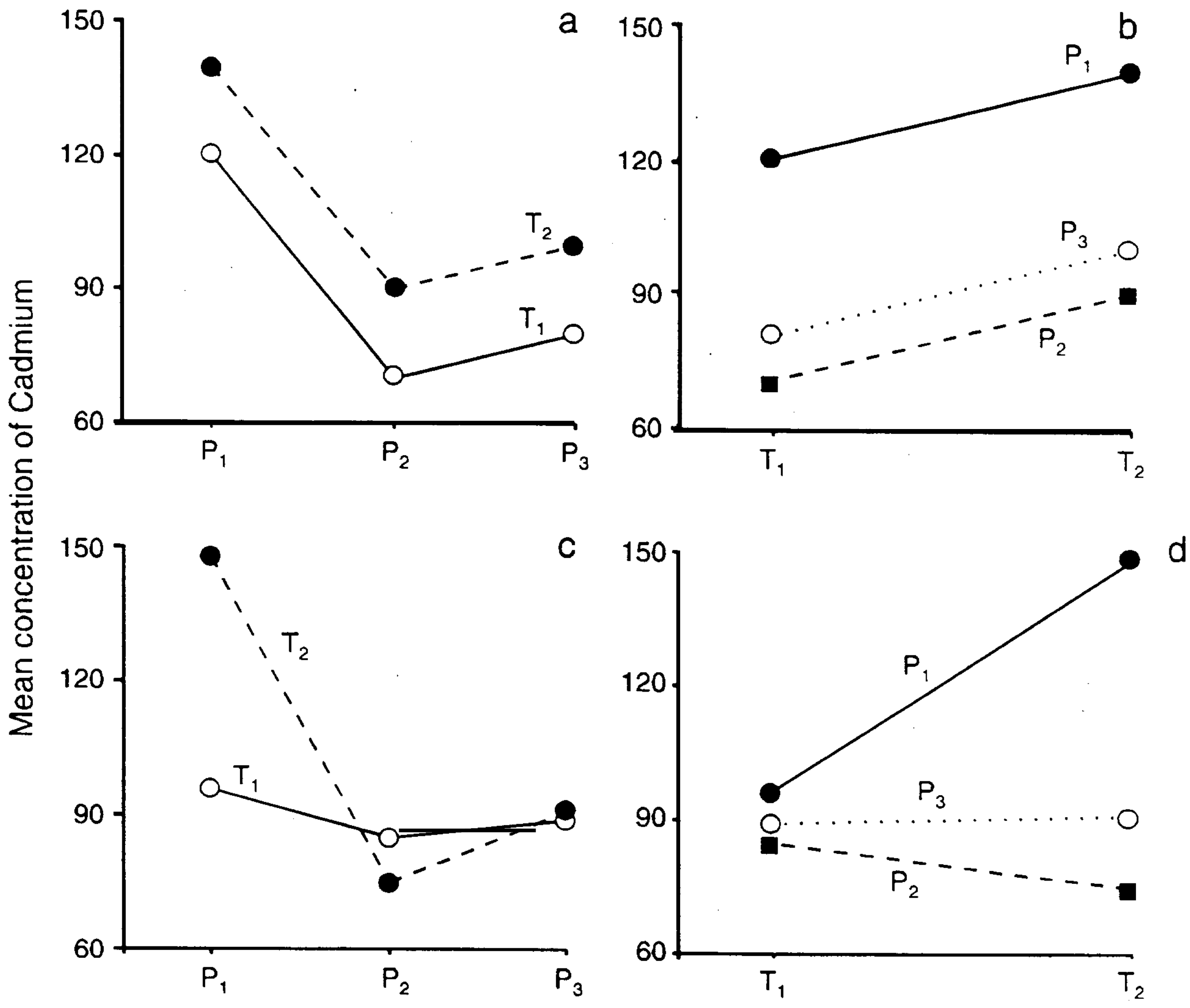

Figure 11. Results of an analysis of variance of two fixed orthogonal experimental factors P and T. P is 3 Places (*a*= 3) and T is 2 Times of sampling, before and after an oil-spill (*b*= 2). The oil-spill is in the first Place (P_1)There were *n* replicates in each of the *pt* combinations of experimental treatments. In (a) and (b), there is no interaction between *P* and *T*. The differences from one level of *P* (one place) to another are equal for each time of sampling. The differences between the two times are identical for all three places. The graphs are parallel. (a) shows the trends of time for each place. (b) shows the same data as differences among places for each time.. In (c) and (d), there is marked interaction between *P* and *T*. There is virtually no change in cadmium concentration from Time 1 to Time 2 in Places 2 and 3. In contrast, there is a major increase in *P*1. The differences between T_1 and T_2 *are* not the same at each level of *P*. The differences from P_1 to P_2, P_1 to P_3, P_2 to P_3 *are* not the same for T_1 and T_2. (c) and (d) are plotted like (a) and (b), respectively.

Essentially, for each level of one factor, a multiple comparison is done of the means of [he other factor. The Residual Mean Square is used to calculate the standard errors for the means, exactly as in the case when there is one factor. An example is given in Table 15. The outcome is clear. There was significantly greater concentration of contaminants with increasing depths, but this was a greater trend in winter. Furthermore, there was no difference between summer and winter at the shallowest depth (10 cm), but seasonal pattern at greater depths (15 and 20 cm). The multiple comparisons have revealed a perfectly interpretable pattern of differences which explain why there is an interaction.

Table 15

Multiple comparisons of a two-factor experiment with interactions. Factor.4 is three depths in the mud; Factor *B* is two times of year. These were chosen to test the hypotheses that concentration of contaminants would vary with depth and from summer to winter. Both factors are fixed,

a) Data are concentrations (*n* = 5 replicates at each depth in each season).

Depth (cm)		10	15	20
Summer	20	0.12	0.11	0.24
Winter	60	0.17	0.47	0.65

b) Analysis of variance

Source of variation	Sum of squares	Df	Mean square	F-ratio	
Depths	0.450	2	0.225		
Seasons	0.560	1	0.560		
D x S	0.190	2	0.095	13.58	*P* <0.01
Residual	0.168	24	0.007		
Total	1.369	29			

$$\text{Standard error for means} = \sqrt{0.007/5} = 0.032$$

c) Student-Newman-Keuls tests; * indicates $P < 0.05$.

Effect of depth in each season

	Depths in Summer			Depths in Winter			
Rank	1	2	3	1	2	3	
Mean	0.11	0.12	0.24	0.17	0.47	0.65	*D*
Difference	$^{3-1}$ 0.13*			$^{3-1}$ 0.48*			0.132
	$^{2-1}$ 0.01	$^{3-2}$ 0.12*		$^{2-1}$ 0.30*	$^{3-2}$ 0.18*		0.109

Seasonal differences at each depth

Season at	Depth 10		Depth 15		Depth 20		
Rank	1	2	1	2	1	2	*D*
Mean	0.12	0.17	0.11	0.47	0.24	0.65	
	$^{2-1}$ 0.05		$^{2-1}$ 0.36*		$^{2-1}$ 0.41*		0.109

8.3.2 When There Is No Significant Interaction

When there is no interaction, multiple comparisons are done on each significant main effect, The means of each factor can, however, be averaged over all the levels of the other factor. The lack of interaction confirms that the differences among levels of one factor are independent of the other factor, so the levels of the other factor can be averaged together.

An example is shown in Table 16. This is a similar experiment to that described in Table 15, but in [his case there is no interaction. There are three depths, but this time there are four seasons. The standard error for the means of each depth is calculated in exactly the same way as before. It is the square root of the mean square of the residual (i.e. within samples) divided by the size of sample from which each mean is calculated. The means are, however, now calculated from samples of 5 replicates in each of 4 seasons of - a total of 20 replicates. For the different seasons, each mean is calculated from replicates (5 in each of 3 depths)." Thus, the calculations of standard errors are as in Table 16.

There is a significant increase in mean concentration when depth is 20 compared with 15 and 10, which do not differ. Independently, there is greater mean concentration in all seasons except Spring and it increased from Summer to Autumn and Autumn to Winter.

8.4 MORE COMPLEX DESIGNS

These designs can be extended to more factors (which lead to more types of interesting and informative interactions). They can involve combinations of fixed and random orthogonal factors and combinations ***of'*** nested scales of temporal or spatial sampling. The general procedures have been described in many texts (Crowder and Hand 1990, Wirier et al. 1991). There is no space here to increase the cover of different designs. Instead, the theory covered so far can be used to consider aspects of practical environmental sampling.

9. LOGICAL AND HISTORICAL BACKGROUND TO ENVIRONMENTAL SAMPLING

Collecting data to detect the presence or measure the size of an environmental impact requires careful design of the sampling programme. Consider the usual situation where there is one disturbed site - one area where a jetty or a power-plant or an outfall or an aquaculture facility will be built. In this area, it is suggested that the environmental disturbance will cause changes in the numbers of a species of worm in the sediments (or will alter the concentration of copper in the blood of fish, or will change the amount of detoxifying enzymes in their tissues, or any other univariate measure). So, the hypothesis being proposed is that the mean value of the chosen measure will be different after the environmental disturbance (after the jetty or power-plant or outfall is built) compared with the values before.

Table 16

Multiple comparisons in a two-factor experiment with no interactions. The sampling has a = 3 depths and b = 4 seasons of the year; n = 5 replicates in each treatment.

a) Mean concentrations of contaminants

Depth		10	15	20	Mean
Season	Spring	0.12	0.14	0.38	0.21
	Summer	0.14	0.17	0.37	0.23
	Autumn	0.28	0.30	0.54	0.37
	Winter	0.43	0.46	0.67	0.52
	Mean	0.24	0.27	0.49	0.33

b) Analysis of variance

Source of variation	Sum of squares	Df	Mean square	F	
Depths	0.463	2	0.231	10.51	$P < 0.01$
Seasons	0.903	3	0.301	13.69	$P < 0.01$
D × S	0.083	6	0.014	0.63	$P > 0.05$
Residual	1.056	48	0.022		
Total	2.505	59			

$$\text{Standard error for mean at each depth} = \sqrt{0.022/20} = 0.033$$

$$\text{Standard error for mean at each season} = \sqrt{0.022/15} = 0.039$$

c) SNK tests; * indicates $P < 0.05$

Depth	Rank	1	2	3	
	Mean	0.24	0.27	0.49	D
		$^{3-1}$ 0.25*			0.11
		$^{2-1}$ 0.03	$^{3-2}$ 0.22*		0.09

Season	Rank	1	2	3	4	
	Mean	0.21	0.23	0.37	0.52	D
		$^{4-1}$ 0.31*				0.15
		$^{3-1}$ 0.16*	$^{4-2}$ 0.29*			0.13
		$^{2-1}$ 0.02	$^{3-2}$ 0.14*	$^{4-3}$ 0.15*		0.11

9.1 SIMPLE BEFORE/AFTER CONTRAST .

The simplest design that could be used [o detect a change in mean abundance of worms before and after some environmental disturbance is a single sample taken in the site before and a single sample taken after [he potential disturbance (Fig. 12). This is widely used in response to certain obvious accidental incidents of potential impact. such as oil spills, if some prior information was available. If subsequent sampling reveals differences, these are attributed to the oil-spill. Obviously, there may be no relationship between the observed event and the change in numbers of an organism. The change may have been due to any cause that happened at the same time as the observed human activity. For example, there may have been a seasonal change in numbers of worms or, while power-plant was being built, a decline in their numbers along the entire coast-line. There are no controls in time or space to demonstrate whether such changes are not widespread without there being an incident of pollution or any other human action in that site.

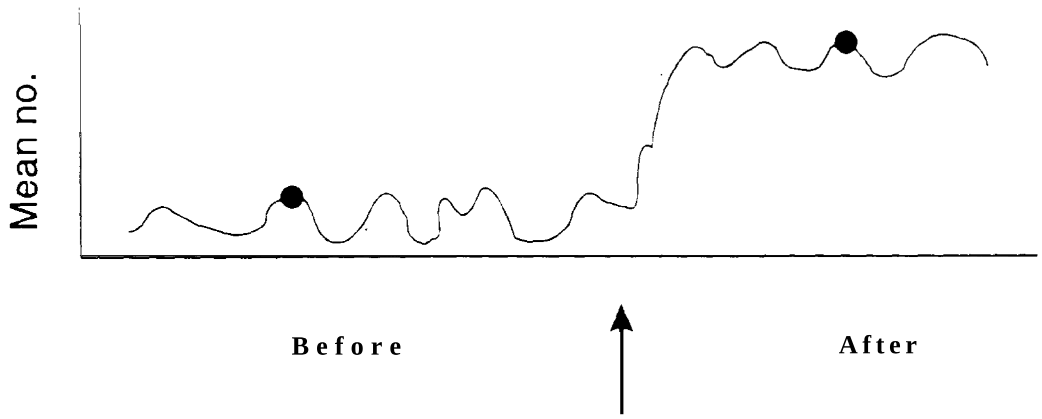

Figure 12. A single sample in one location before and after an impact (at the time of the arrow) is sampled to detect the impact.

It is impossible to make any decision about an environmental impact because any change found from before to after the disturbance may be due to any cause - not just the obvious disturbance.

9.2 MULTIPLE SAMPLES, BEFORE AND AFTER A DISTURBANCE

The other problem with this sort of environmental sampling is that a single sample taken at one time before a disturbance will not represent the variable being measured if it fluctuates in value. So, if numbers of worms naturally change from one month (or season) to the next, you would expect a change from a “before” sample to one taken some months later, after an outfall is built.

Many people have suggested that the solution is to take repeated, randomly timed samples at the disturbed location before and after the development (or accident) that might cause environmental change (Figure 13), Such a design is, again, only able to detect that a change in mean numbers of the sampled population has occurred coincidentally with the onset of the potential environmental disturbance. Data of this sort cannot indicate that there is some relationship between the change in numbers and the putative disturbance (Underwood and Peterson 1988, Underwood 1989, 199 1). The data are "pseudoreplicated" (Hurlbert 1984): any change in numbers may be due to any number of causes (including that identified as a disturbance by man). It is remarkable how often such sets of data are used (e.g. Bachelet 1986, Butanone and Moore 1986, Dauvin and Ibanez 1986, Lopez-Jamar et al. 1986) or are even recommended (Lundalv et al. 1986) in environmental work. This is not an appropriate procedure.

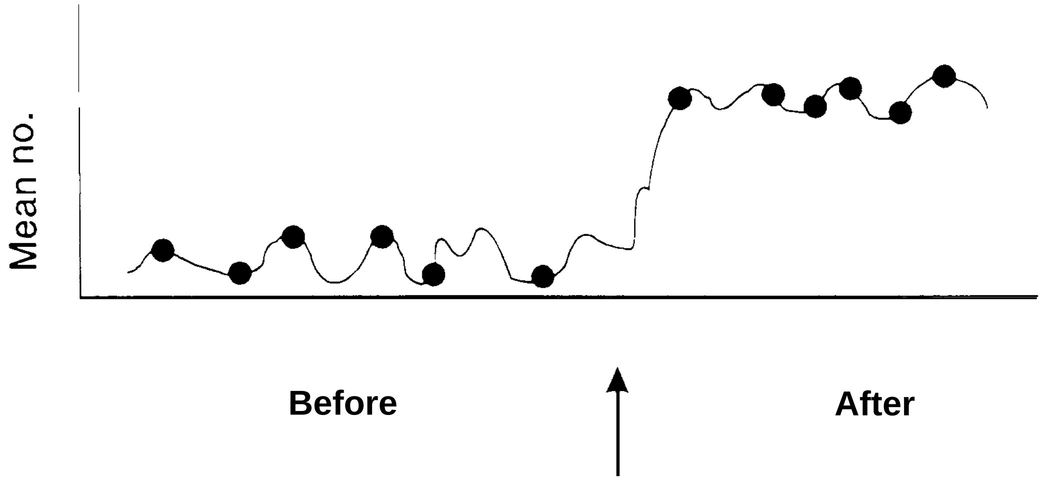

Figure 13. Random samples are taken at different times in one location before and again after an impact (at the time of the arrow) in an attempt to detect the impact.

Stewart-Oaten et al. (1986) identified the great advantages of taking the samples at random intervals of time, rather than on some fixed schedule. They suggested, taking samples at randomly placed intervals of time will tend to ensure that no cyclic differences unforeseen by the sampler will influence the magnitude of the difference before and after the onset of a potential environmental disturbance. Taking samples at regular, as opposed to randomly chosen, intervals means that temporal variance will not be estimated accurately and that the magnitude of an impact may be over- or underestimated. There are such serious consequences that regular temporal sampling should be avoided (see the more extensive discussions in Green 1979, Stewart-Oaten et al. 1986, Underwood 1991).

9.3 COMPARISON OF A DISTURBED TO A SINGLE CONTROL (OR REFERENCE) LOCATION

Sometimes, instead of taking data before "the disturbance and comparing the measurements with what happened afterwards, environmental impacts are investigated by comparing the disturbed area with a single so-called control or reference area.

Most natural populations, measures of biochemistry, physiology, etc., vary from place to place when there are no disturbances caused by humans. It is therefore impossible to say that a difference from one place (where [here has been a disturbance) to another (where there has not) has been caused by the disturbance. So, even though the number of worms in the disturbed place may be smaller than the numbers in the control place, that might be the natural situation and' not caused by the arrival of an outfall or a power-plant or a harbour.

9.4 THE BACI (BEFORE/AFTER, CONTROL/IMPACT) PROCEDURE

Green (1979) suggested that a so-called BACI design would generally be the most useful for detecting environmental change. His design (Figure 14) involved a single sample taken in each of two locations. One is the potentially disturbed location - usually called the “Impact” location - even though no impact is yet known to be occurring. The second location is a similar area to serve as a control that can be sampled in identical fashion and independently of any change in the first location. Ideally, the two locations are sampled at the same time before and again after the developments.

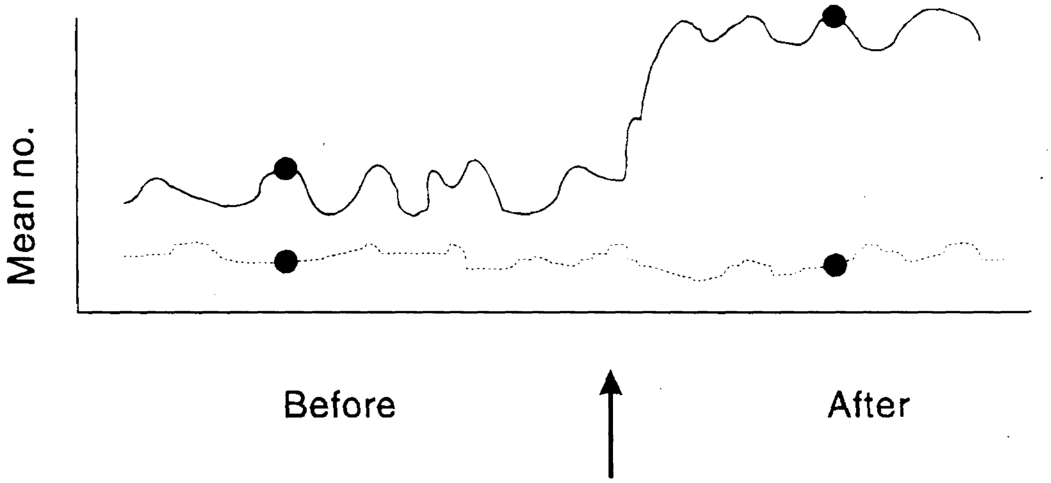

Figure 14. BACI design as proposed by Green (1979), with a single sample before and after the impact (at the time of the arrow) in each of a control (dashed line) and the possibly impacted location (solid line).

The hypothesis being examined is that the difference from before to after a disturbance in the ‘Impact’ location will not be the same as the natural change from before to after in the control location,

As a result of having two locations, provided that the usual procedures are done to ensure that samples are representative, unbiased, independent and so forth (Snedecor and Cochran 1967, Dixon and Massey 1969, Green 1979; see Underwood 1981 for marine examples), the data are straightforward to analyse. The best test is a two-factor analysis of variance (provided its assumptions can be met). The existence of some environmental impact would be shown by a significant statistical interaction between the two sources of variation (Control versus Impact locations and Before versus After times of sampling) (Table 17b). Such an interaction would indicate that the magnitude of difference from before to after in the mean numbers of the sampled population in the Control location is not the same as the difference from before to after in the means in the Impact location. This is a logical procedure for detecting differences in the amount of change in two populations from one time to another. It is not, however, correct to conclude that any interaction is **caused** by the human activity in one location that did not occur in the other, control location. There may be intrinsic changes due to any number of processes that cause changes in mean abundance in a population. It is not uncommon, particularly in marine examples, for populations in two locations to diverge or converge through time without there being an environmental impact caused by man at some particular time in one of the two locations. This is a common feature of natural populations (Underwood 1989). Thus, use of a single control location is not adequate to associate the presence of the statistical interaction with the impact of the specified activity in the impact location (see the related discussion in Bernstein and Zalinski 1983), Several control locations should be used (see later).

Table 17

BACI: A single time of sampling at two locations, one Control and one potentially Impacted.

Sources of variation		Degrees of freedom	*F* ratio versus	Degrees of freedom
Before vs. After	=B	1		
Locations: Control versus Impact	=L	1		
Interaction BxL		1	Residual	1,4 (*n*- 1)
Residual		4 (*n*- 1)		
Total		4 *n*- 1		

9.5 BACI: REPEATED BEFORE/AFTER SAMPLING AT CONTROL AND IMPACT LOCATIONS

One of the most widely used designs is a modification of Green’s (1979) BACI design and was foreshadowed by him as an extension of that method. It was firs: formally analysed by Bernstein and Zalinski (1983) and later discussed in detail by Stewart-Oaten et al (1986). The design involves sampling the location that is planned to be affected by some development and a single control location. Each location is sampled several times, at random, before and then after the start of the potential disturbance. The two locations are sampled at the same times (i.e. times of sampling are orthogonal to the two locations) and there are similar (ideally, the same) numbers of samples taken before and after the planned development. As discussed in detail by Stewart-Oaten et al. (1986), times of sampling are random and thus form a random, nested source of variation in the data (Underwood 1981, Wirier et al. 1991). The times of sampling are nested in either the period before or the period after the potential impact starts.

The rationale behind the design is that the replicated sampling before the development gives an indication of the patterns of differences, over several periods of potential change of numbers of organisms, between the two locations. If the development in a location causes a change in the mean abundance of the sampled organisms, there will be a different magnitude of difference between the two locutions after it starts from that prevailing before. By having several replicated times of sampling, it is possible to control for some random elements of difference between the two locations, in contrast to Green's (1979) design. It is no longer possible for the two locations to differ after the possible impact simply because, at a single time of sampling, they happen to have a different pattern of difference from that before. To be detected, an impact must cause a sustained pattern of difference.

Both Bernstein and Zalinski (1983) and Stewart-Oaten et al. (1986) considered that the best statistical analysis was to calculate the differences between the mean abundances in the two locations for each time of sampling and to analyse the difference from before to after in the means of these differences (Figure 15).

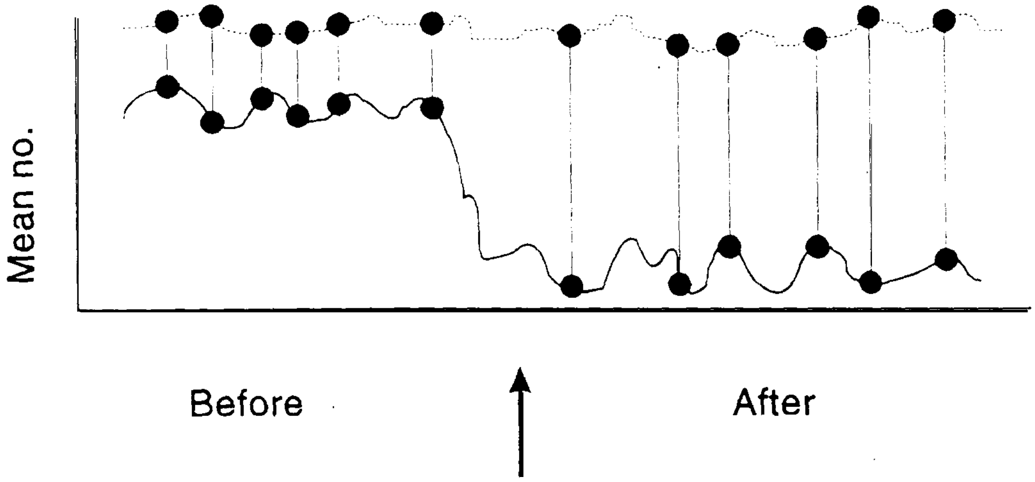

Figure 15. BACI according to Bernstein and Zalinski (1983) and Steward-Oaten et al. (1986); differences between mean abundances in the control and potentially impacted locations are calculated for random times of sampling before and after the disturbance begins (at the time of the arrow). Vertical lines indicate differences between the two locations.

Unfortunately, despite the assertions of authors such as Stewart-Oaten et al. (1986), the repeated-sampling BACI design is still impossible to interpret. Discovery of a different pattern of difference between the two locations from before to after the disturbance does not demonstrate that the human activity **caused** the impact. There is still no reason to expect two locations to have the same time-course of changes in mean abundance of a given species (or any other variable). What is needed is a set of control locations and then to demonstrate that the pattern of change in mean abundance of the population being sampled from before to after the onset of the human disturbance is greater in the Impact location than occurs on average in the Control locations (Underwood 1992).

9.6 BEYOND BACI USING SEVERAL CONTROL LOCATIONS

The solution to the problems identified above is to use several control locations. There cannot be several disturbed locations - you only intend to build one outfall or one power-plant or one prawn-farm. You can, however, examine several undisturbed locations.

There should be a series of locations, randomly chosen out of a set of possible locations that have similar features to those of the location where the development is being proposed. The only constraint on random choice of locations is that the one planned to be disturbed must be included in the sample. This is not nearly as difficult as it seems; the locations do not have to be identical, in any sense of the word. They simply have to follow the normal requirements that they come from a population of apparently similar locations. This is the basis of all random sampling theory. Of course, the locations should be independently arranged. For most marine habitats, locations can be sufficiently widely spaced that they are not correlated by processes of recruitment or disturbances.

The logic of the design is that an impact in one location should cause the mean abundance of animals there to change more than expected on average in undisturbed locations. Abundances in the control, undisturbed locations will continue to vary in time, independently of one another. On average, however, they can be expected to continue as before. Note the important difference from BACI design. Here, the individual control locations may differ and may change significantly from one another. On average, however, the set of controls should continue to show average behaviour. Impacts are those disturbances that cause mean abundance in a location to change more than is found on average.

Formally, the hypothesis that there is going to be some impact of a proposed development is a statement that after the development has started, there will be some difference between the potentially impacted location from the average of that in the controls. This hypothesis is in its simplest (and most primitive) form and will be developed later. It is, however, instructive to examine this as opposed to the simple notion that a contrast between one control and a potentially impacted location is all that is required. The analysis is developed starting in Table 18. What is required is one planned (a priori) orthogonal contrast (Scheffé 1959, Wirier et al. 1991). This contrast is between the potentially impacted (identified as I in Table 18) and the control locations. Differences among the control locutions are not of intrinsic interest, but a real environmental impact requires there to be greater difference between the impacted and the control locations than there is among the controls.

Examination of Figure 16 will illustrate the principle of the procedure. First, the abundances of some organism may show any pattern of difference among the locations before a possible impact. These will presumably continue whether or not an impact occurs. Thus, differences among locations are to be expected. Second, as discussed above, there is no reason to presume that the differences among locations are constant through’ time. Thus, there will be statistical interactions between locations and time of sampling. In this particular design (Figure 16), sampling is done twice, once before and once after the disturbance. It is reasonable, therefore, to expect that the differences among locations are not the same before and after. There should be an interaction, identified as B x L in Table 18.

Table 18.

Asymmetrical sampling design to detect environmental impact; 1 Locations are sampled, each with n random, independent replicates, once Before and once again After a putative impact starts in one Location (“Impact”); there are (f- 1) Control Locations; Locations represent a random factor.

Sources of variation		Degrees of freedom		
Before vs. After	=B	1		
Among Locations	=L	*(l-* 1)		
Impact vs. Controls[a]	=I		1	
Among Controls[a]	=C		*(l-2)*	
BxL		*(l-* 1)		
BxI[a,b]			1	
BxC[a]		*(l-2)*		*F-* ratios vs. Residual
Residual		*2l(n-* 1)		
Total		*2ln-* 1		

[a] Repartitioned sources of variation,

[b] Impact can be detected by the F-ratio Mean Square B x I/Mean Square B x C. If B x C is not significant (the system is non-interactive), impact can be detected by the F-ratio Mean Square B x I/Mean Square Residual (see text for details).

If, however, there really is some effect on the population in the disturbed location (as in Figure 16), the difference between that location and others before the impact occurs should not be of the same magnitude (or, for some locations, the same direction, as in Figure 16). In the example illustrated, mean abundance is reduced by the impact, causing a larger negative difference from that in the location with the greatest abundance, a smaller positive difference from that in the location with the smallest abundance and a smaller (as opposed to larger) abundance than that in the other location (Figure 16).

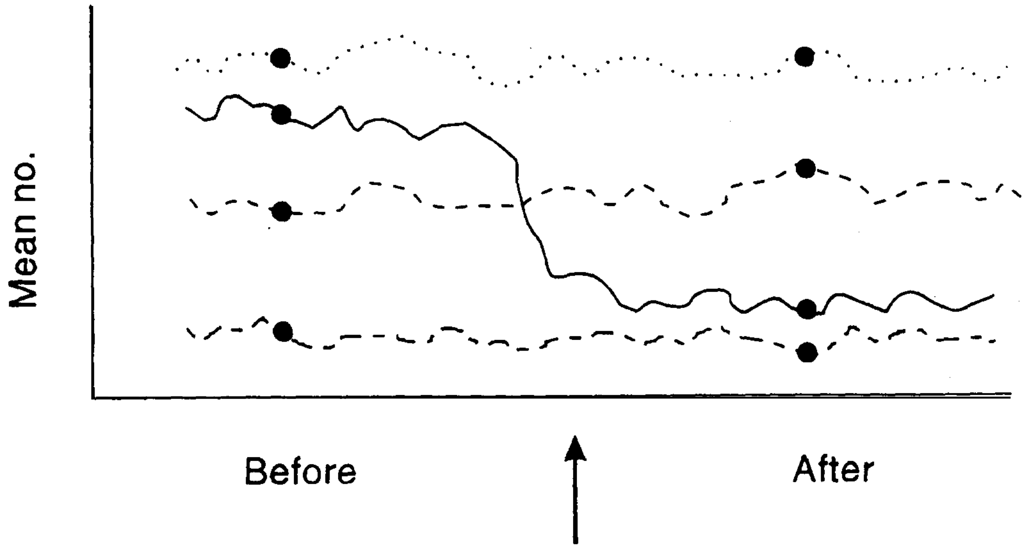

Figure 16. Sampling for assessment of environmental impact with three control locations and a single impacted location (indicated by the arrow at the right). Sampling is done at a single time before and after the impact begins.

In an analysis of such a situation (Table 18), the impact would be detected as an interaction between time of sampling (Before versus After; B in Table 18) and the difference between the Impacted and Control locations (I in Table 18). This interaction is identified as B X I in Table 18.

The locations are in the sampling programme in two sets of groups. One is the set of control locations, which are randomly chosen and represent a random factor in the design. The disturbed and possibly impacted location is the sole (i.e. unreplicated) member of the other group. Thus, the appropriate model for the analysis is a form of replicated repeated measures design (e.g. Hand and Taylor 1987, Crowder and Hand 1990, Wirier et al. 1971), except that one group (the “group” of a single Impact location) is spatially unreplicated (although there are replicates within the location).

It is, of course, still possible that this interaction could be significant because of some differential fluctuations in the abundances of organisms in the different locations that have nothing to do with the disturbance, In this case, there should generally be interactions between the time of sampling (B in Table 18) and the differences among the Control locutions (C in Table 18), regardless of what happens in the disturbed Location. This would be detectable as a significant B x C interaction in Table 18.

An environmental impact on the abundance of a population in some location should be defined as a disturbance that causes more temporal change in the population than is usually the case in similar populations in other similar locations where no such disturbance occurs. Formally, the hypothesis is that the interactions in time between the possibly impacted and a set of control locations (B x I in Table 18) should be different from the naturally occurring interactions in time among the Control locations (B x C in Table 18).

If disturbance does cause a small impact, such that the magnitude of change in the population is small, an ecologically realistic interpretation is that the fluctuation in the impacted population is within the bounds of what occurs naturally elsewhere. It is within the resilience of natural populations (e.g. Underwood 1989) and therefore no cause for concern. The interaction from before to after the disturbance between the putatively impacted and the average of the control locations (B x I in Table 18) must be created by an environmental disturbance. The appropriate tests is an F-ratio of the Mean Square for B x I divided by the Mean Square for B x c.

If, however, it can be shown that there is no natural interaction among the Control locations from before to after the disturbance (i.e. B x C is itself not significant when tested against the Residual Mean Square), an impact would be detected by more powerful F-ratio of the Mean Square for B x I divided by the Residual mean Square. This test makes some assumptions about the data, which are described in detail in Underwood (1992). They are not important here. If the test for B x I were significant, it would indicate the presence of an unnatural change associated with the impacted location.

If the test were not significant, it would not mean that the disturbance has had no effect. Rather, it indicates that whatever effects have occurred are too small to cause any serious change in the population being examined. This design still suffers from the problem that there is no temporal replication and therefore chance differences, not related to environmental impacts would also be detected and interpreted as being an impact.

9.7 BEYOND BACI WITH SEVERAL SAMPLES BEFORE AND AFTER AN IMPACT

An appropriate analysis would have replicated sampling at several times before the development and several times after, as in the Bernstein and Zalinski (1983) and Stewart-Oaten et al. (1986) procedure, but in the potentially impacted and in the replicated control locations. From these data, it is possible to ascertain whether there is an interaction between the difference between the impacted and control sites through time. This is much more likely to be caused by the disturbance itself than any simpler analysis. The procedure for analysis is illustrated in Table 19. There is one contrast between the potentially impacted and the other locations.

Sampling is done in all locations at the same times. but these times are randomly chosen, This ensures that the changes in each of the locations are not measured in a temporally corrected way (see the discussion in Stewart-Oaten et al. (1986) on this point). Although it is often considered that sampling for environmental impact should be done on some seasonal, or other humanly chosen time course, this is not a good idea. Even though there may be strict seasonal cycles in the abundance of some species, unless they are absolutely the same from year to year, sampling from one 3-monthly period to another may not represent the seasonal pattern at all (see examples in Stewart-Oaten et al. 1986). It is better to sample randomly through time, but not at too widely different time periods so that it is reasonable to determine temporal variance without having imposed different sorts of temporal patterns into the data (Stewart-Oaten et al. 1986), The best designs would have several temporal scales of sampling (Underwood 1991), see later.

This design now has several key features of importance. First, there is spatial replication. Second, there is temporal replication before and after the proposed development. Finally, there is the possibility of detecting impacts as an actual change in abundance in only the impacted location, which would appear as a difference between after and before in that location only. This design also allows formal tests on the amount of interaction in the impacted versus the control locations as demonstrated in Table 19.

Sampling is done at several (t) independent times at random intervals before and after the possible impact begins. Independence through time may be difficult to achieve. All locations are sampled at the same times; time of sampling and locations are fully orthogonal factors in the sampling design. If sampling cannot be done at the same time, a different analysis must be used.

Analysis for environmental disturbance then depends on the time-scale of the effects of the impact. In the simplest case, consider the situation when there is no short-term temporal interaction among control locations after the disturbance starts (i.e. T(Aft) x C is not significant in Table 19). There is no difference in the temporal pattern from one control location to another. A disturbance that affects the impacted location so that it differs from the controls can then be detected using the F-ratios of T(Aft) x I divided by the Residual, as indicated in Table 19 (footnote a).

Table 19

Asymmetrical sampling designs to detect environmental impact; 1 Locations are sampled, each with *n* random, independent replicates, at each oft Times Before (= Bef) and *t* Times After (= Aft) a putative impact starts in one Location ("Impact"); there are *(l-* 1) Control Locations; Locations represent a random factor and every Location is sampled at the same time; Times represent a random factor nested in each of Before or After.

Sources of variation		Degrees of freedom	
Before vs. After	=B	1	
Among Times (Before or After)	=T(B)	2 *(t-* 1)	
Among Locations	=L	*(l-* 1)	
Impact vs. Controls	=1	1	
Among Controls	=C	*(l-* 2)	
B x L		*(l-* 1)	
B x I$^{1\text{-}c,d}$		1	
B x C$^{1\text{-}c}$		*(l-* 2)	*F-* ratios vs. Residual
T(B) x L^{1}		2(*t-* 1) *(l-* 1)	
T(Bef) x L^{1}		*(t-* 1 *) (l-* 1)	
T(Bef) x I^{a,1,b}		(t-l)	
T(Bef) x C^{1,b}		*(t* -1 *(l-* 2)	
T(Aft) x L^{1}		*(t-* 1 *) (1-* 1)	
T(Aft) x I$^{1\text{-}a}$		(t-1)	
T(Aft) xC$^{1\text{-}a}$		(t-1) *(l-* 2)	*F-* ratios vs. Residual
Residual		2*lt*(*n*-1)	
Total		*21tn-* 1	

a If T(Aft) x C is not significant, impact can be detected by:
(i) F=Mean Square T(Aft) x L/Mean Square Residual is significant
(ii) 2-tailed F=Mean Square T(Aft) x I/Mean Square T(Bef) x I is not significant
(iii) 2-tailed F=Mean Square T(Aft) x C Mean Square T(Bef) x C is not significant

b If T(Aft) x C is significant, impact can be detected by (see text for details):
(i) F=Mean Square T(Aft) x I/Mean Square T(Aft) x C is significant;
(ii) 2-tailed F=Mean Square T(Aft) x I/Mean Square T(Bef) x I is significant;
(iii) 2-tailed F=Mean Square T(Aft) x C/Mean Square T(Bef) x C is not significant

c If and only if there are no short-term temporal interactions in a, b and B x C is not significant, impact can be detected by F=Mean Square B x L/Mean Square Residual.

d If and only if there are no short-term temporal interactions in a,b, but B x C is significant, impact can be detected by F=Mean Square B x I/Mean Square B x C

Often, however, fluctuations in abundances from time to time vary significantly from location to location even when there is no human disturbance. The interaction among control locations from time to time of sampling before and after the disturbance will then be significant. This will occur when relatively large changes in numbers of populations occur out of phase in different places. A disturbance affecting the population in an unnatural manner must cause a changed pattern of temporal interaction after it begins. The impact must cause and altered pattern of differences between the mean abundance in the impacted and those in the control locations. This is obvious in Figure 17; consider at, each time of sampling the differences in abundances between the impacted location and the location with the greatest abundance. These differences are altered by the impact.

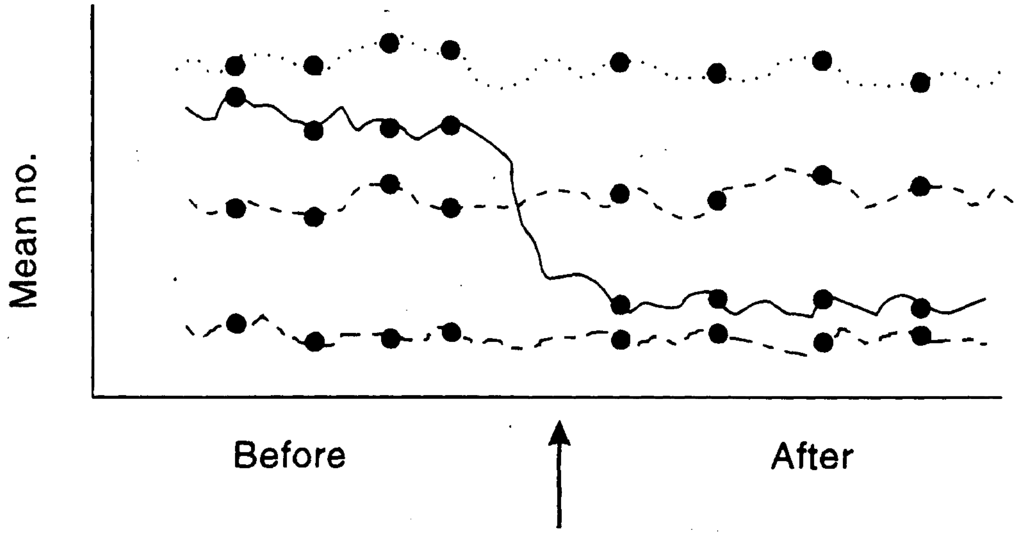

Figure 17. Sampling for assessment of environmental impact with three control locations and a single impacted location (indicated by the arrow at the right). There are 4 times of sampling before and again after the impact starts.

So, there should be a difference between the interaction between times of sampling and the differences between impacted and control locations that occur after the impact starts (T(Aft) x I in Table 19) and this pattern of interaction among the control locations after the impact starts (T(Aft) x C in Table 19): Furthermore, the pattern of interaction (T(Aft) x I in Table 19) should no longer be the same as occurred before the impact started (T(Bef) x I in Table 19). Both of these can be tested by *F*- ratios identified as footnote [b] in Table 19.

Finally, any change in the interaction found after the disturbance between the impacted and control locations might be due to general changes taking place at the same time as the disturbance. To demonstrate that the change is associated with the disturbance and not part of a general change there should be no change in the temporal interactions among control locations after (T(Aft) x C) compared with before (T(Bef) x C) the disturbance. This is tested in Table 19, footnote b.

Two of these tests ((ii) and (iii) in footnote b in Table 19) are 2-tailed, unlike traditional l-tailed *F*- ratios in analysis of variance. This is because interactions among times of sampling may be increased or decreased from before to after a disturbance starts.

In biological terms, these tests detect whether an impact has caused some change to the population in one location making it vary from time to time differently from the temporal pattern found on average in control locations (T(Aft) x I will differ from T(Aft) x C). The second test determines whether the disturbance makes the population different from before in the time-course of its abundance (T(Aft) x I will differ from T(Bef) x I).

There is also the possibility that an environmental impact occurs in a system which is not interacting temporally (T(Aft) x C in Table 19 is not significant) and causes an impact of greater duration.

Thus, after the disturbance, there is no increased interaction from time to time between the impacted and control locations. Instead, the impact is a continued change in the population in one location after the impact starts (Figure 17), causing a larger difference in the abundance after versus before in the impacted location than would be found in the control locations. Where there is no shorter-term temporal interaction (T(Aft) x C, T(Aft) x I are not significant in the tests described previously), the more sustained impact (affecting B x I in Table 19) can also be tested.

First, consider what to do when the control locations vary together from before to after the disturbance (i.e. B x C is not significant when tested against the Residual in Table 19). Under these circumstances, an impact should cause the difference between impacted and control locations after the disturbance to differ from that before. Thus, B x I will be significant as identified in footnote [c] in Table 19.

If, however, there is a general change in the mean abundances of populations in all locations that is coincident with the beginning of some disturbance, B x C would also be significant. Under these circumstances, an impact must cause a larger B x I interaction than that measured as B x C. An impact must make some pattern of temporal difference in the affected location that is unusual compared to those that naturally occur elsewhere. This can be tested by the F-ratio of the Mean Squares for B x I and B x C (in footnote b of Table 19). This test for B x I would not be very powerful, unless there were many control locations (making 1 large in Table 19).

9.8 BEYOND BACI WITH CAREFUL DESIGNS OF TEMPORAL SAMPLING

It is very important to sample at time-scales relevant to the problem and hypotheses. Suppose that numbers of worms in sediment fluctuate very much from day to day and week to week in response to tidal or lunar processes. If they are sampled every three months, the differences from one sample to the next are not interpretable. There may be differences due to the fact that one sample was taken during a period when there were large numbers, but small numbers would have been found a week later.

Therefore, appropriate temporal patterns of sampling must be identified so that sampling is at time-scales consistent with the rates of turn-over and the life-cycles of the organisms being sampled. This is obviously not always a simple matter (as discussed in Underwood 1991).

The choice of when to sample is often arbitrary. It is commonly set by previous practice ('monthly sampling is routine in this field'), logistic constraints ('we cannot get to the site more often') or some sort of conceptual model ('seasonal variation is important'). These may be good guide-lines, but the timing and frequency of sampling should really be determined by logical thought about the processes operating and the organisms being examined.

The problem of sampling at an inappropriate time-scale is illustrated in Figure 18. The actual number of organisms per unit area fluctuates from time to time without any seasonal trend (Figure 18). Seasonal sampling (over fifteen months) with replicated sampling units is done once in each season, resulting in estimates of mean abundance as in Figure 18. These sampled data show a marked seasonal trend, which is erroneous. The error occurred because shorter-term fluctuations were not sampled. There is no replication within seasons and therefore comparisons among seasons are confounded (mixed up) with shorter-term changes. The replicated samples at each time in each season do not estimate temporal fluctuations in each season. Increased frequency of sampling at shorter intervals would solve the problem, but it would be better to design the sampling programme to include replicated temporal samples in each season.

This is illustrated in Figure 18c, where seasonal sampling is done by having several (three here) independent, randomly chosen, replicate temporal samples in each season. The average of these (asterisks in Figure 18c) now correctly reveals no seasonal trend. The choice of two hierarchically arranged frequencies of sampling removes the confounding. Thus, replicated sampling at smaller time-scales is always necessary to demonstrate that temporal trends are not caused by shorter-term fluctuations (see Underwood 1993 for further examples).

There are also specific, biological reasons for choosing more than one time-scale at which to sample. The appropriate scale may not be known, there may be processes operating at different time-scales (say seasonal episodes of recruitment on a background of tidally driven, fortnightly changes in supplies of planktonic food; Underwood 1991).

The analysis is made more difficult by the requirement that data from each time of sampling be independent of each other, or lacking in serial correlation (see the discussions in Cochran 1947, Eisenhart 1947, Swihart and Slade 1985, 1986, Stewart-Oaten et al. 1986). This is a complex topic, but generally requires that data are taken at sufficiently long intervals that there is no influence from one time to another. Data will not be independently sampled if samples are taken at such short intervals at the same individuals or cohort of individuals are repeatedly counted. Under these circumstances, the numbers at one time will be the numbers at a previous time minus mortality and emigration. Thus, subsequent counts must be smaller or equal to the previous numbers and are not independent. If mortality or emigration are themselves density-dependent, there are obviously going to be correlations between the numbers at one time and those found previously. Independence of the data is not a requirement unique to these analyses, it is a major assumption of most statistical procedures (see, for example, Underwood 1981). It is however, an assumption often overlooked in biological samples and experiments (e.g. Cochran 1947, Eisenhart 1947, Gurevitch and Chester 1986, Swihart and Slade 1986).

However many times the populations are to be sampled and at whatever interval, the times chosen should be random, not regular (see the discussion in Stewart-Oaten et al. 1986). In that way, the times of sampling are extremely unlikely to coincide with some cyclic pattern and thereby cause problems of interpretation of the data. So, if sampling is to be at the two time-scales of a few days apart and at periods of 6 months, there should be several randomly chosen days of sampling every 6 months. For the most powerful capacity to detect an impact, the locations should be all sampled at the same time and at all the times.

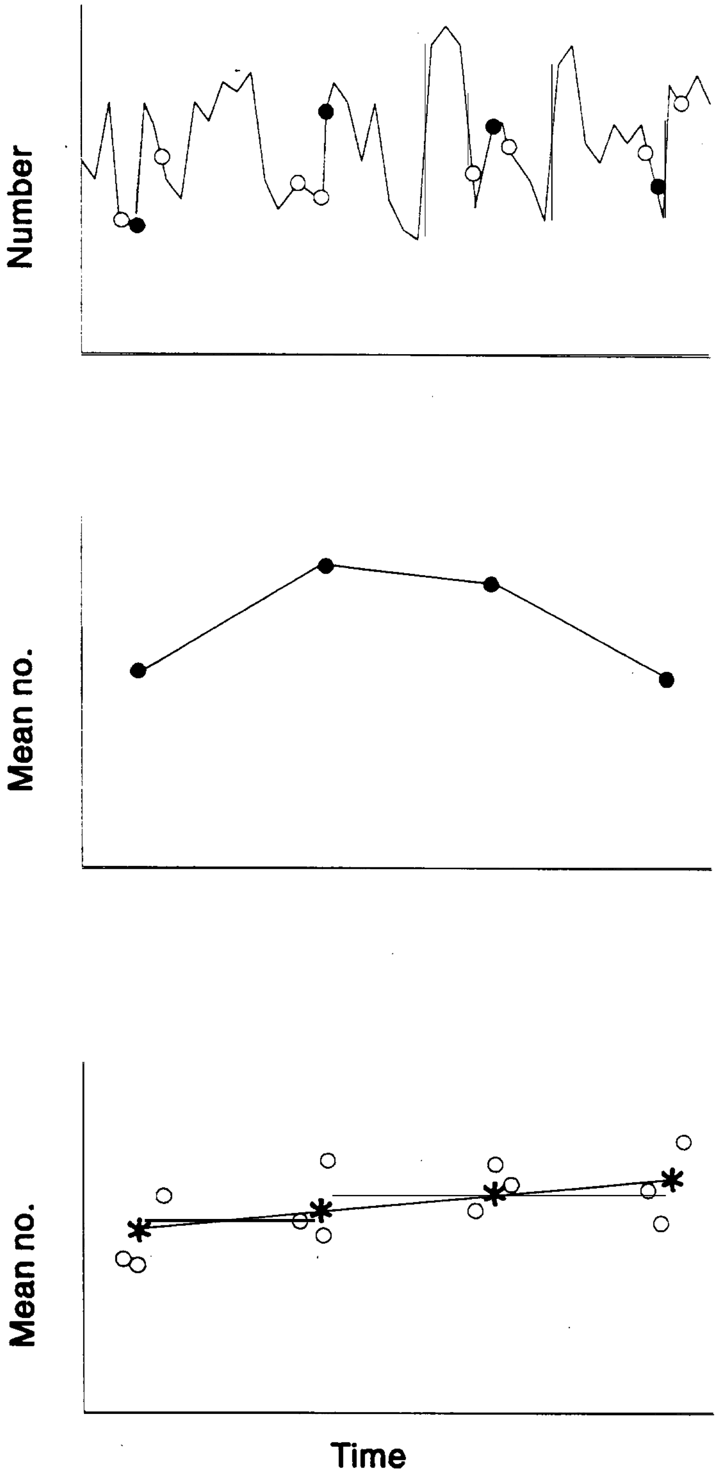

Figure 18. Sampling to detect temporal trends in abundance of a population. (a) There is no seasonal trend in a fluctuating population. Samples are taken at one time in each season (●) or at three times (●and ○). (b) Apparent seasonal trend when only a single time is sampled in each season. (c) Better representation when three replicated times are sampled in each season(○), with mean abundance (*) showing no seasonal trend.

10. PROBLEMS WHEN THE SPATIAL SCALE OF AN IMPACT IS NOT KNOWN

The extent or spread of a possible environmental impact is often not knowable before it occurs. Designing a sampling programme to detect it is therefore very difficult. Consider the following case of impact due to dredging for airport runway in a coastal bay. There is concern that dredging will stir up contaminants in sediments and these will cause reductions in numbers of invertebrates in nearlby patches of seagrass. The engineers hope the potential effects of dredging are very localised. They will contain silt to prevent it spreading. So, they establish sampling sites in the area to be dredged and in other randomly-chosen seagrass beds around the bay (as in Figure 19). A comparison of the patterns of mean abundance of invertebrates in the disturbed site to the temporal patterns of change in the control sites, from before to after the disturbance should reveal the impact (Underwood 1992).

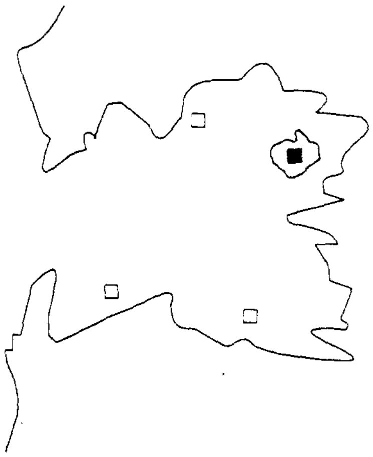

Figure 19. Sampling designed to detect small-scale environmental impact at a disturbed site (solid square) in a bay, Three control sites are randomly scattered around the edge of the bay. The impact (shaded) is predicted to be small.

Environmental scientists are, however, worried about what will happen if dredging stirs up contaminants bound to very fine sediments which are washed over the whole bay (as in Figure 20). Now, there can be no detection of an impact if it occurs - it will affect all of the control sites and no difference will be seen in any analysis.

The solution is to have replicate control sites in other bays (or elsewhere) that are sufficiently far away or isolated from the disturbed site to guarantee no possible impact (as in Figure 21). These will then serve to identify changes at the scale of entire bays (or other spatially large units). The design must therefore include spatial sampling at appropriate nested (or hierarchical) scales. The analysis of the data must be sufficient to cope with two levels of spatial scale (Underwood 1992, 1993).

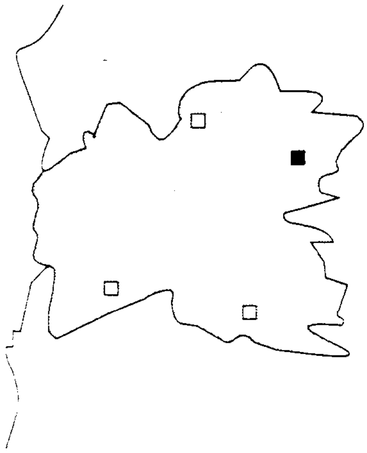

Figure 20. Sampling design as in Figure 4. The impact (shaded) turns out to be large and also affects the entire bay.

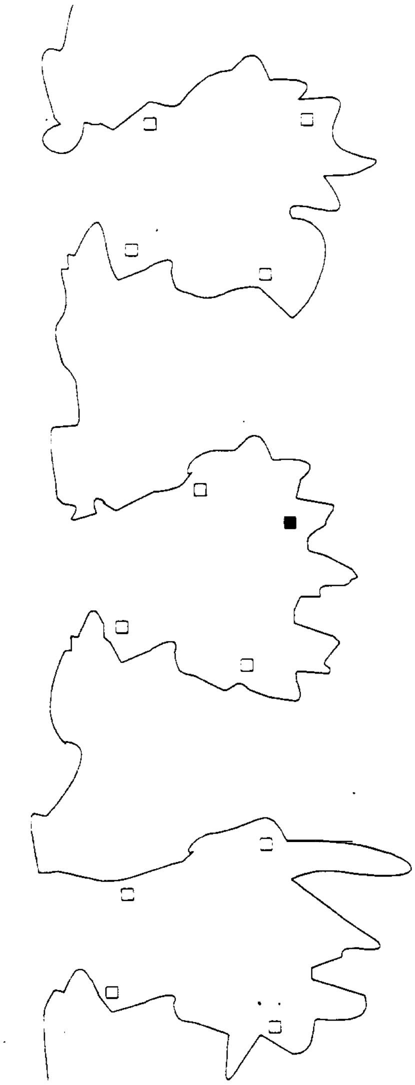

Figure 21. Sampling designed to detect small- or large-scale environmental impact in a bay. If it is small, the impacted site will differ from the control sites in the same bay. If it is large, the impacted bay will differ from control bays.

Again, statistical consultants would be better able to advise on. appropriate sampling if biologists and ecologists (and regulatory agencies) put much more thought into the spatial scale of potential impacts. This requires much more effort and discussion than has often been the case.

10.1 MAKING SURE SPATIAL REPLICATES REALLY ARE PROPER REPLICATES

Another problem that must be thought about is that sampling usually requires measurements to be made in several places as replicates to allow proper estimation of spatial variation. Consider the case of an oil-spill onto patches of coral reef or (seagrass beds). The oil hits some patches and not others. Suppose we want to measure the effects by measuring the cover of coral (or the numbers of shoots of seagrass) in quadrats in oiled and unoiled areas. The hypothesis being examined is that there will be less cover (or fewer shoots) in the oiled areas.

Consider what happens if we measure in replicate quadrats arranged in space in oiled and unoiled locations like those in Figure 22. Any interpretation of an analysis for differences between the two locations at any one time is completely impossible. All the analyses can do is to reveal a difference between the two areas. Any difference may be due to a disturbance in one of them as proposed in the hypothesis. But this assumes that they were identical in the cover of coral (or number of shoots) before the oil-spill and would have been similar after it were it not for the disturbance. Of course, the cover of coral (or number of shoots) varies greatly from place to place where they have not had an oil-spill. So, any difference found in sampling may be natural and not due to the oil.

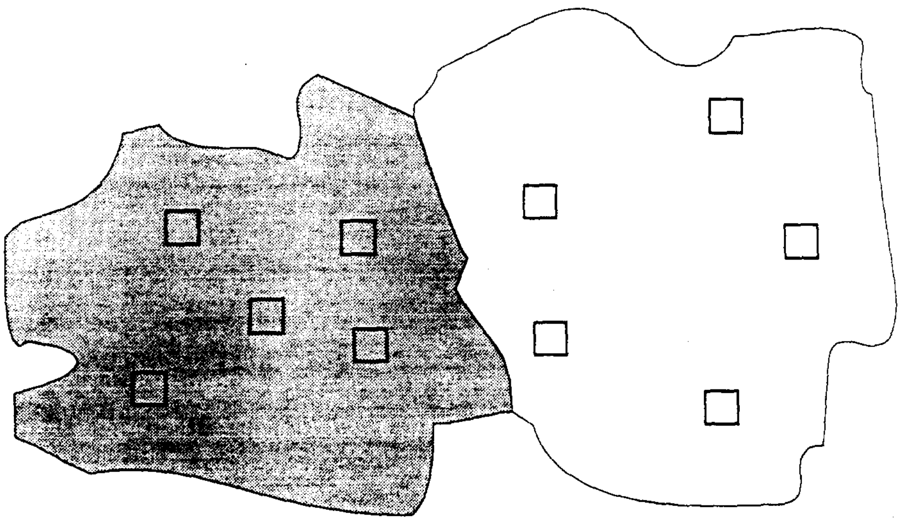

Figure 22. Confounded (unreplicated) sampling to detect environmental impact in an accidentally oiled (shaded) area in comparison to a control (open) area. Each is sampled with five quadrats.

What is needed is a sampling design with properly interspersed, independent, representative patches of undisturbed (i.e. control or reference) sites and potentially impacted sites. Such an arrangement is illustrated in Figure 23.

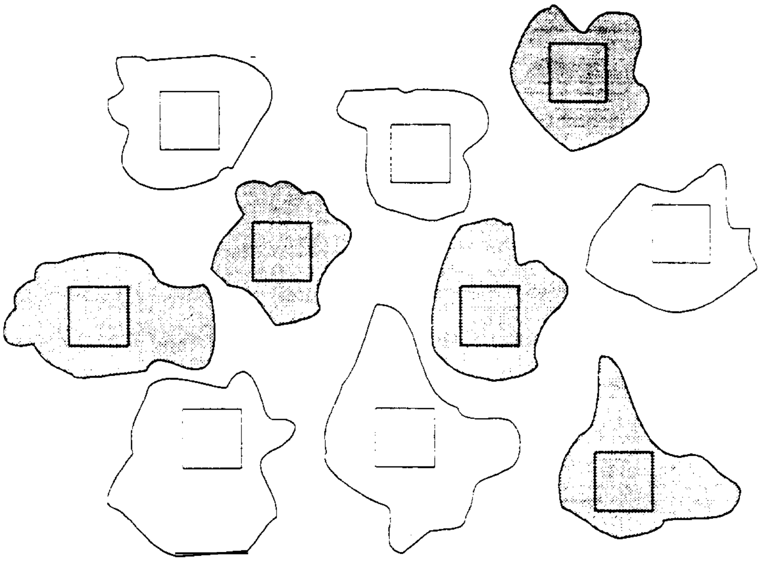

Figure 23. Properly replicated sampling of two habitats (shaded and open) each sampled in five independent patches.

Unfortunately, making the recommendation that ecologists or environmental scientists must have replicated sites does not guarantee that they get them. Sometimes, sampling is done in independently scattered control and possibly impacted sites, but these are really still in a single location. This is shown in Figure 24. The oil-spill is widespread and it is possible to have several sites within it and to have similarly distant sites in a nearby control. There are still going to be all sorts of reasons why cover of coral may have been different even if there had been no oil.

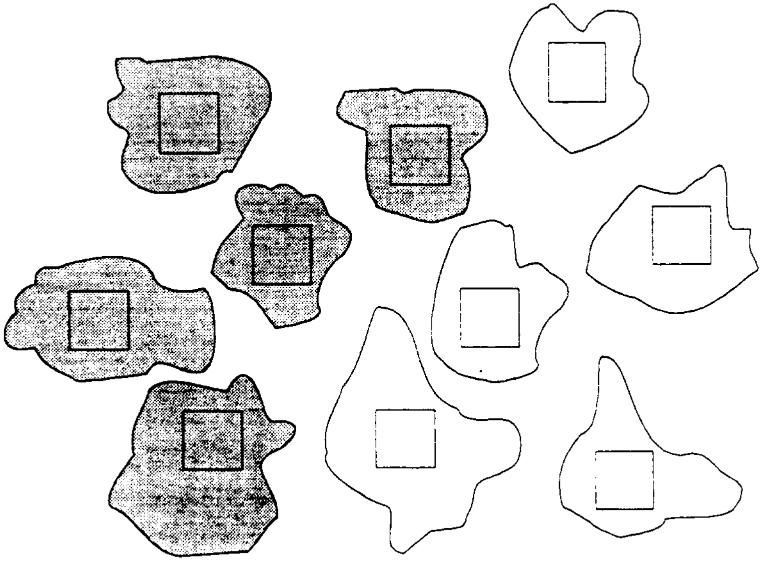

Figure 24. Confounding caused in sampling to detect environmental impact in accidentally oiled (shaded) areas in comparison to control (open) areas. Five patches of each type were sampled, but the oiled patches were not interspersed with control ones.

This sort of design keeps creeping into environmental studies. For example, in the case of an oil-spill on a coral reef, there is a large disturbance over many hectares. Patches of the area are undisturbed areas (e.g. Green 1993), A random sample of these and of sites with oil can then be compared. Any difference in, the average cover of live coral could then be blamed on the oil and it be defined as an impact.

Such an inference is still confounded by potential intrinsic differences between oiled and unoiled sites. Although they are spatially scattered (as in Figure 22), they may really be like those in Figure 24. The reason for there being no oil in some patches may be. (and probably is) due to them having different physical characteristics. Some patches may be shallower (or deeper) or have different hydrology or different original structural composition Any of these could cause these to be the patches that received no oil. Any of these variables probably also influences the cover of coral or any other variable you may want to measure. Consequently, the comparisons are not interpretable unless it can be demonstrated that there were no differences in biological variables in these patches before the oil arrived. Large-scale environmental disturbances are nearly always accidental and unplanned, making it impossible to design sampling programmes before they occur (but see the suggestions in Underwood 1989, 1993, 1994b). So, careful thought is needed about where to take samples.

11. TRANSFORMATION TO LOGARITHMS

For analysis of variance of many types of environmental sampling, it may be advantageous to transform the data to logarithms before doing the analyses. There are two situations where this is an appropriate procedure.

First, one of the assumptions for analysis of variance is that variances of each sample are the same (see discussion of homogeneity of variances earlier). Often when variances are not all the same, transformation to logarithms will help, particularly where data come from very skewed frequency distributions. These usually occur when sampling rates, ratios, concentrations and are therefore quite common when sampling environmental variables.

The second case where logarithms of data are appropriate is where an environmental impact can be expected to cause a **proportional** change in the variable being measured. For example, the accidental release of a toxic chemical into an area of estuarine mudflat may kill an average 20% of the worms there. If there were about 50 per m^{2}before the spill, after the pollution there would be a reduction to 40 per m^2, a difference of 10 per m^2. On the other hand, if there had originally been 80 per m^2, there would now be 64 per m^2, a difference of 16. In log scale, both impacts would be the same. In the first case, there would be a reduction from log$_e$(50) to 1og$_e$ (40) = 3.91-3.69, a difference of 0.22. In the second case, there would again be a reduction of 0.22, from log$_e$ (80) = 4.38 to log$_e$(64)= 4.16. So, log scales are useful for making different impacts comparable where the variables being measured are of different sizes. Logs are the appropriate scale when impact are **multiplicative,** that is they alter the variable of a proportion. Logarithms are often useful where variances of samples are very different (see further discussion in Green 1979, Underwood 1981, Wirier et al. 1991).

Table 20

Simulated data to illustrate procedures for calculation of sums of squares in asymmetrial analyses of a possible environmental impact that might affect the density of an organism.

Location	Putatively impacted			Control 1			Control 2		
Before the Disturbance									
Time of sampling									
1	59	51	45	46	40	32	39	32	25
2	51	44	37	55	47	41	31	38	45
3	41	47	55	43	36	29	23	30	37
4	57	50	43	36	44	51	39	29	23
After the Disturbance									
Time of Sampling									
5	38	44	52	31	38	45	42	35	28
6	52	44	37	51	43	37	38	31	24
7	60	52	46	30	37	44	41	34	27
8	53	46	39	40	34	26	21	27	35

Data were simulated for three locations, one of which is planned to have a disturbance and the other two arc controls. Data arc for four times of sampling before and four times after the disturbance; at each time of sampling there arc n= 3 replicate counts of organisms in each location. The data arc for a situation where there is no disturbance.

12. A WORKED EXAMPLE

To show how to analyse a set of data for assessing an environmental impact, a worked example is provided in Table 21. The data are from an area where a new outfall pipe has been built from a factory, The local Environmental Protection Authority wants to test the hypothesis that the chemicals have altered the numbers of crustaceans in core-samples of mud. The EPA has data for 4 times of sampling before the outfall was built (Times 1- 4) and 4 times after it started operating (Times 5- 8). At each time, 3 replicate cores were taken at the location where the outfall was planned and at 2 similar areas of mudflat (control locations 1 and *2). So,* altogether, there are 3 locations, 8 times of sampling and 3 replicates in each of the 24 samples.

The data must be transformed to natural logarithms before analysis because we expect that if there is an impact, it will kill some proportion of the worms. There are two ways to do the analysis. First, you can construct all the appropriate formulae from texts such as Wirier et al. (1991) or Crowder and Hand (1990) and use the right formulae in your statistical package. This can be very complicated.

Alternatively and much more simply, do several simpler analyses. First, analyse all the data. Then re-analyse the data omitting the possibly impacted location. Each of these analyses is very straightforward and can be done as a fully orthogonal analysis Before versus After, Locations and a nested factor, Times of sampling within Before or After. Such a model is easily handled by many statistical packages, The results” of such analyses can be manipulated by addition and subtraction to calculate the quantities required,

The worked example is presented in Tables 21 and 22 to demonstrate the procedures. The data are in Table 20 to allow checking of other programmes and readers’ own calculations.

Four analyses are done. First (Table 21a), all the data are analysed in the full model, with all times, locations, before and after the disturbance, considering the locations as one set (i.e. not divided into control and impacted groups). Second, as in Table 21b, the data are analysed as a three-factor analysis of all the data from control locations, without the potentially impacted location. Then, in order to handle the nested component, that is Times within Before or After, analyse the set of all the data from before the disturbance (analysis c in Table 21) and finally analyse the data from before, but only the Control locations (analysis 21d). From these four analyses, the entire asyrnmetrical analysis as required can be calculated by subtractions and additions of components. The algebra is all indicated in Tables 21 and 22.

This worked example is only for calculations, but if you wanted to practise interpreting analyses, in this case there is a major impact which reduced all the densities of worms in the impacted location after the disturbance. So, there was a significant interaction from before to after in the difference between the Impacted and Control locations (B x I in Table 22). The Mean Square of 3.388 would turn out to be significant if you did the appropriate test. In this example, there is no short-term temporal interaction, so all the terms (T(Bef) x I, T(Bef) x C, T(Aft) x I, T(Aft) x C in Table are not significant (their Mean Squares are in the range 0.011 to 0.056 compared with the Residual which is 0.031).

Table 21

Calculations of analyses of environmental impact for data for one putatively impacted and two control locations, each sampled at the same 4 randomly scattered times before and again after a planned disturbance. Data have been transformed to $\log_e(x)$.

Analysis		a		b		c		d	
Source of variation		Sum of squares	d.f.	Sum of squares	d.f.	Sum of squares	d.f.	Sum of squares	d.f.
Before *vs* After	=B	2.371	1	0.038	l				
Times (B)	=T(B)	0.219	6	0.377	6	0.101	3	0.223	3
Locations	=L	al 1.078	2	b^{1}0.557	I	1.015	2	0.381	I
BxL		a^{2}3.404	2	b^{2}0.016	l				
T(B)xL		a^{3}0.378	1 2	b^{3}0.104	6	c^{1}0.200	6	d^{1}0.033	3
Residual		1.744	48	1.368	32	0.843	24	0.672	16
Total		9.185	71	2.460	47	2.159	35	1.309	23

There were 3 replicate samples at each time in each location. Data arc in Table 20. Four analyses are needed to complele all calculations: a, all data; b, excluding the disturbed ('Impact') location; c, all locations before the disturbance; d, before the disturbance, excluding the impact location. See Table 22 for use of superscripts.

Table 22

Analysis of an impacted and two control locations sampled four times before and four times after a disturbance
(data arc in Table 20 and preliminary analyses are in Table 21).

Source of Variation		Sums of squares	d.f.	Mean squares	Calculated from
Before vs After	=B	2.371	1	2.371	a
Times(B)	=T(B)	0.219	6	0.037	a
Locations	=L	1.068	2	0.534	al
Impact vs Controls	=I	0.51	1	0.511	al -bl
Among Controls	= C	0.557	1	0.557	bl
BxL		3.404	2	1.702	a2
BxI		3.388	1	3.388	a2-b2
BxC		0.016	1	0.016	b2
T(B)xL		0.378	12	0.032	a3
T(Bef)xL		0.200	6		c1
T(Bef)xL		0.167	3	0.056	cl -dl
T(Bef)xC		0.033	3	0.011	dl
T(Aft)xL		0.178	6		a3-c1
T(Aft)xL		0.107	3	0.036	a3-c1-b3+d1
T(Aft)xC		0.071	3	0.031	b3-dl
Residual		1.744	48	0.036	a
Total		9.185	71		a

The origin of the calculated sums of squares is as in Table 21 (see superscripts in Table 2 I).

13. POWER TO DETECT ENVIRONMENTAL DISTURBANCES

One of the big problems in most environmental research is how to calculate, in advance, the power of any statistical analysis to detect the biological effects. The topic has been discussed before (e.g. Bernstein and Zalinski 1983, Fairweather 1991, Green 1979, Peterman 1990, Underwood 1993. 1994a), The logic and background will only be summarized here. Consider a simple example of samples taken at one time in, say, five locations to examine whether there is a difference in mean abundance of animals (or mean concentration of pesticides, or any other variable) from one location to another. The data can be analysed and a statistical test may reveal some difference. In other words, if the statistical test is significant at some chosen probability, often $P = 0.05$, differences are presumed to exist. Either there is a real difference in the means among locations and the result is correct, or there is no important difference among locations. Alternatively, the samples do not represent the mean abundance very well and therefore have erroneously caused detection of an apparent difference. The probability of this error occurring by chance when the null hypothesis is true (and there really are no differences in mean abundance among locations) is $P = 0.05$ (as chosen by the experimenter). This is the probability of Type I (or α) error, the probability of rejecting a null hypothesis even though it is true. This probability is determined by the investigator, a priori by choice of the value of “significance” (i.e. the probability that would cause rejection of the null hypothesis).

There is, however, the opposite potential error of failing to reject a null hypothesis when it is, in fact, false. If there are differences in mean abundances from one location to another, but few samples are taken and/or there is much variation in abundance from one sample to another within locations, it is unlikely that the statistical test used would detect the difference (e.g. Andrew and Mapstone 1987, Wirier et al, 1991). The probability of this error occurring is the probability of Type 11 (or β) error (Figure 25). This probability is determined by how much difference really exists among the populations (the effect size), how many samples are taken, the intrinsic variability in the quantity being measured and the chosen level of significance (α, the probability of Type I error).

The power of a statistical procedure is its capacity to reject, when appropriate, a null hypothesis. In the case considered, this is the capacity to discover differences among populations when they exist, that is to reject an incorrect null hypothesis. Common-sense dictates that the power of a test is therefore the probability that rejection of the null hypothesis will occur properly when it is false, which is the complement of a Type II error. Therefore, power is defined as (1-Probability of Type H error), as in Figure 25.

It is no more clear what is a desirable probability of Type 11 error than is the case for adoption of the conventional 0.05 for probability of Type I error. At first sight, however, it would appear reasonable under most circumstances to assume that the probability of Type I and Type H errors should be the same. Thus, if an error occurs, it has an equal chance of being in each direction (which is one very good reason for never drawing conclusions from single experiments, or samples, or locations, even if they are very powerful). Thus, if *a* is $P = 0.05$, then β should be 0.05 and power will be 0.95.

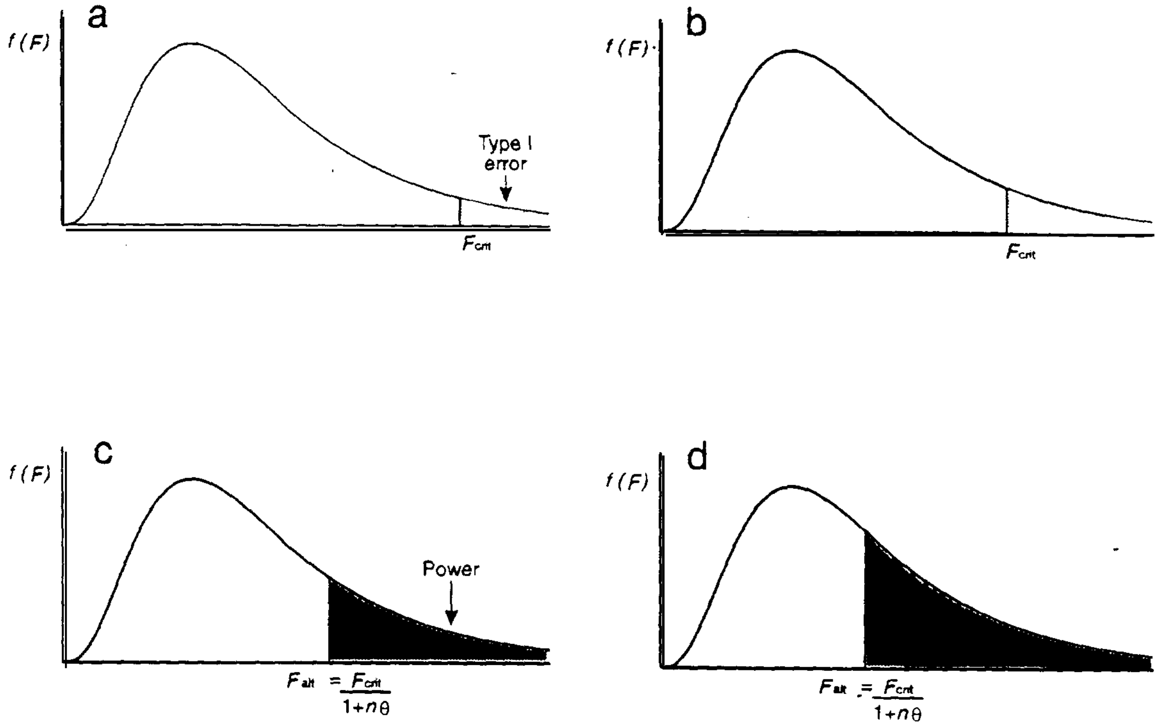

Figure 25. Diagram to illustrate the nature of calculations of power to detect environmental impacts using analysis of variance of random effects (for complete details, see Underwood, 1993). In (a) and (b), the relevant null hypothesis is true and there is no impact. The distribution of the test statistic, *F*, is then as shown. In (a), the probability of Type I error (detecting an environmental impact when none is present) is set at 0.05, as is traditional. In (b), the probability is 0.10 (note the larger shaded area representing Type I error in (b)). F_{crit} *is* the critical value of *F* - if from your samples, the calculation of *F* gives a value larger than F_{crit}, YOU would declare there to be an environmental impact. A value smaller than F_{crit} would be evidence of no impact. In (c) and (d), there is an impact of a particular size, causing a particular amount of difference between impacted and control locations. The size of this difference is related to θ, the "effect size". Note that the value of *F* that would cause you to conclude there is an impact is identified as F_{alt} (i.e. for this specific value of θ). So, if you get a value of *F* larger than F_{alt}, you would conclude (correctly) that there is an impact. The power of the testis now the shaded region representing all values of *F* that would correctly cause you to identify the impact. Note that power is increased by increasing a, the probability of type I error (note the difference between (c) and (d) corresponding to the difference in a in (a) and (b)). Also, an increase in the effect size - the size of impact to be detected by sampling - increases power, because θ is in the divisor for determining F_{alt}. Finally, the size of samples *(n)* influences power. If you have larger samples (i.e. larger *n)*, you are more likely to detect the impact (because, again, *n* is in the divisor for determining $F\sim_{ll}$; larger values of n lead to smaller values of F_{alt} and therefore greater probability of getting a value of *F* 1arger than F_{alt} in your statistical test).

This must, however, be modified to consider the consequences of the two types of mistake. If a potential pollutant is being evaluated to determine its effects on abundance of some organism and several polluted and several control sites are available, either error may occur in a test for differences between the average abundance in polluted and that in control sites. A Type II error would imply that the pollutant was not causing any difference to the abundances when, in fact, there were effects but they had not been detected. Thus, management of the problem would declare there to be no problem, despite the future deleterious consequences. In contrast, a Type I error would imply that the pollutant was harmful, even though there is actually no real difference in average abundance. Management would require prevention of discharge, or reduction of the concentration or rate of discharge, or rehabilitation of the habitat,

Each decision has a cost. The latter is calculable in dollars. There are direct and many indirect costs of regulation of discharge, clean-up or redesign of an alternative disposal (let alone tines for non-compliance, etc.). The former (Type II error) is more difficult to assess, but is a “social” cost of altered environment. loss of diversity of fauna, etc.

Because it is difficult to know the real cost of an environmental change, it is often appropriate to weight the probabilities of the two types of error, so that one is more likely to make a Type I error. Thus, a should he chosen to be large. This is known as the precautionary principle (Bodansky 1991). Making a large will make it likely that if an error does occur, it will be to reject a true null hypothesis and thus to conclude that there is an impact, when, in fact, there is no real effect of a disturbance. The assumption in this is that the “cost” of failing to detect an impact is difficult to predict, is usually non-recoverable and may seriously reduce quality of life, exploitation of species, etc. The costs of detecting something that is not actually an impact are often known - absolutely in the case of projects that involve technological or industrial development, where the potential environmental change is brought about as a result of some proposed “profit” or “usefulness”. If there is doubt that a statistically significant pattern has really identified an environmental impact, more work, more sampling (and better designed sampling to increase the power of tests) can be planned to resolve the problem. Meanwhile, increased regulation, damage control and attempts at rehabilitation can be implemented. If it then turns out that there was impact (i.e. the result was found to be a Type I error), no harm has been done to the biological system after all. The extra delays and costs can be argued as necessary to improve understanding of the system (see the discussion in Fairweather 1991, Peterman 1990).

So, it is important to attempt to evaluate any sampling design to determine how powerful it will be to detect a possible impact of any predetermined magnitude. Many environmental studies use analyses of variance of mean values of variables. These involve statistical tests called F-ratios. In general, the F-ratios used are so-called variance component tests (Wirier et al. 1991). The calculation of power is straightforward. The relationships between expected values F when a particular null hypothesis is true and when a specified alternative is true are illustrated in Figure 25, for two different chosen levels of probability of Type I error (α).

What is evident is that increasing α must increase power. This is common sense and is well-known. Note in Figure 25 that increasing α from 0.05 (Figure 25a) to 0.10 (Figure 25b) greatly increases the power to detect differences among means (compare Figure 25c with 25d). Second, power will be greater where the “effect size” (the magnitude of difference that should occur if an impact exists) is larger.

So, in Figure 25, θ identifies the “effect size”. If a large difference between disturbed and control locations is expected due to a potential effect of the disturbance, θ will be large and F_{alt} will be small, so there is a good chance that your data will produce a calculated value of F that is larger than F_{alt}. So, if the anticipated impact occurs (θ is, in fact, large), you are likely to reject the null hypothesis and detect the impact. Your test is powerful (i.e. likely to cause you to identify the impact).

Conversely, power to detect a particular effect size is decreased where intrinsic, natural variability (σ_ϵ^2) is large. Where σ_ϵ^2 is large relative to the effect size (θ), $1 + n\theta$ will not be much greater than 1 and therefore Fait will not differ much from F_{crit}. This is because the size of the effect is small relative to the size of natural variability. Under these circumstances, you are not likely to detect the impact and the power of the test is small. Finally, as would be expected from common-sense, power is greater where there are larger samples (either in terms of *n,* the number of replicates per sample or in terms of the number of control locations and/or times sampled). Again, this is obvious in Figure 25. For large *n,* Fait will be small relative to F_{crit} for any size of effect (for any value of θ). So, there is every chance of getting a value of *F* larger than F_{alt},causing you to reject the null hypothesis and find the impact. So, your test is powerful - you are likely to detect an impact if one occurs that is at least as large as proposed in the hypothesis.

Thus, modelling of potential magnitudes and deviations of effects of disturbances is necessary to estimate the power of any planned sampling. The data gathered before the disturbance can be used as a starting basis for the simulation of realistic effects (e.g. Walters 1986, Underwood 1991, 1992, 1994a). Inevitably, some quantitative estimates of intrinsic variation (i.e. in the undisturbed state) from time to time and among replicates are necessary before such simulation is possible. These can be gained directly from sampling before a planned disturbance. They may also be estimable from other studies on similar populations in the same sort of habitat (as discussed in detail in Underwood 1989, 1992). Until structured comparative sampling of many different types of organisms over extensive time periods has been properly described, collated and the spatial and temporal variances interpreted, sampling prior to a planned disturbance remains the only option.

Relatively few environmental studies consider the power of the procedures they use. As a result, there are probably many examples of environmental change that go undetected. We are pre-occupied by a - the error of finding changes or differences when there is none - instead of β - the error of *not* finding differences when they exist.

Calculation of power of sampling programmes is often difficult, requires training and suggests that seeking expert help would be advantageous. Some help with related issues can be obtained in the book by Cohen (1977). The main reason usually given for not attempting to calculate the power of a sampling programme to detect an impact of defined size is the cost of sufficient samples to achieve large power. This is an important consideration, but not, in fact, the most common problem. A bigger problem is to get the client or regulatory agency to define the magnitude of potential impacts so that the effect size (θ) can be specified.

The cost of doing powerful sampling maybe large if small impacts (small θ) or large variability (large σ_ϵ^2) are anticipated, requiring the number of samples *(n)* to be large. On the other hand, sampling with inadequate power to detect the relevant impacts is a wasted expense that achieves nothing. It is important to understand the issues so that cost-effective sampling can be designed that might detect the defined impacts.

14. CONCLUSIONS

These guide-lines have ranged over a variety of topics to span many of the components of an environmental study. Obviously, the topics have not been dealt with in depth. Some suggestions for further thought are offered here. Two activities will increase the worth, validity, logical structure, scientific content and cost-effectiveness of environmental sampling.

14.1 TRAINING

Methods for measurement, the variety of variables or indicators to measure and the techniques for ensuring quality or measurements all change from time to time. Many laboratories provide training in the use of techniques. Acquire all the information available to ensure quality control of measures on chemical, biochemical and other components of a study.

Statistical techniques and procedures for planning and designing sampling programmes are also changing because both are active correct areas of research. Many organizations run workshops and professional courses for environmental scientists. Some of these are published (e.g. Underwood and Fairweather 1993). Try to attend them or arrange one for your area.

In the attached list of references, those most helpful on aspects of design and interpretation of environmental studies are indicated by an asterisk.

14.2 CRITICISM

Critical appraisal of the planning, implementation, analysis and interpretation of studies. is the only known way to progress. Discuss the plans and the designs of studies with a wide range of people. Keep discussing the issues, the procedures, the logic. Do not hide the design of a sampling programme until it is too late to improve it.

Subject every phase of an environmental study to critical expert review. Everyone involved in environmental and ecological work makes mistakes. There are no exceptions. The difference between good and bad environmental studies is, however, usually based on whether or not the scientists or consultants are learning from the mistakes. For far too long, many environmental professionals have been of the view that the scientific content is not a major priority in applied scientific studies of environmental issues. They seem to believe in the relevance of a distinction between “pure” and “applied” science. In fact, if scientific information is to be used in environmental decisions, it had better be very good information. “Applied” science needs better science than do many areas of “pure” science. To increase the scientific content of environmental work requires greater effort to increase the good science - the logic, the structure, the identification of hypotheses, the rigorous planning of sampling, the use of modem analytical techniques. Above all, therefore, improvement requires vigilance and constant criticism.

15. ACKNOWLEDGEMENTS

The production of this contribution was supported by a grant from the Australian Research Council. I am deeply grateful for the huge help provided by Dr M.G. Chapman and by P. Barnes, R. Cawley, J. Cunningham, J. Harris, V. Mathews and J. Winzar in the preparation of the text and figures.

16. LITERATURE CITED

Andrew, N.L. & B.D. Mapstone, 1987. Sampling and the description of spatial pattern in marine ecology. *Ann. Rev. Oceanog. Mar. Biol.,* Vol. 25, pp. 39-90.

Bachelet, G., 1986. Recruitment and year-to-year variability in a population of *Macoma balthica.* In, *Long-term changes in coastal benthic communities,* edited by C. Heip, B.F. Keegan & J.R. Lewis, Junk, Dordrecht, pp. 233-258.

Bernstein, B.B. & J. Zalinski, 1983. An optimum sampling design and power tests for environmental biologists. *J. Environ. Manage.,* Vol. 16, pp. 335-43.

Bodansky, D. 1991. Law, scientific uncertain and the precautionary principle. *Environment,* Vol. 33, pp. 43044.

Box, G. E. P., 1953. Non-normality and tests on variances. *Biometrika,* Vol. 40, pp. 318-335.

Buchanan, J.B. & J.J. Moore, 1986. Long-term studies at a benthic station off the coast of Northumberland. In, *Long-term changes in coastal benthic communities,* edited by C. Heip, B.F. Keegan & J.R. Lewis, Junk, Dordrecht, pp. 121-127.

Clarke, K. R., 1993. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Eco.,* Vol. 18, pp. 117-143.

Cochran, W. G., 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics,* Vol. 3, pp. 22-38.

Cochran, W.G. & G. Cox, 1957. *Experimental designs, second edition.* Wiley, New York.

Crowder, M.J. & D.J. Hand, 1990. *Analysis of repeated measures.* Chapman and Hall, London

Dauvin, J.C. &F. Ibanez, 1986. Variations à long-terme (1977-1985) du peuplement des sables fins de la Pierre Noire (Baie de Morlaix, Manche Occidental): analyse statistique de l'évolution structural. In, *Long-term changes in coastal benthic communities,* edited by C. Heip, B.F. Keegan & J.R. Lewis, Junk, Dordrecht, pp. 171-186.

Day, R.W. & G.P. Quinn, 1989. Comparisons of treatments after an analysis of variance. *Ecol. Monogr.,* Vol. 59, pp. 433-463.

Einot, I. & K.R. Gabriel 1975. A study of the powers of several methods of multiple comparisons. *J. Am. Stat. Ass.,* Vol. 70, pp. 574-583.

Eisenhart, C., 1947. The assumptions underlying the analysis of variance. *Biometrics,* Vol. 3, pp. 1-21.

Fairweather, P. G., 1991. Statistical power and design requirements for environmental monitoring. *Aust. .J. Mar. Freshwat. Res.,* Vol. *42,* pp. 555-568.

Green, R. H., 1979. *Sampling design and statistical methods for environmental biologists.* Wiley, Chichester.

Green, R. H., 1993. Application of repeated measures designs in environmental impact and monitoring studies. *Aust. J. Ecol.,* Vol. 18, pp. 81-98.

Gurevitch, J. & S.T. Chester, 1986. Analysis of repeated measures experiments. *Ecology,* Vol. *67,* pp. 251-255.

Hand, D.J. & C.C. Taylor, 1987. *Multivariate analysis of variance and repeated measures.* Chapman and Hall, London, 212.

Hartley, H.O., 1950. The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika,* Vol. 37, pp. 308-312.

Hurlbert, S. J., 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.,* Vol. 54, pp. 187-211.

Kennelly, S.J. & A.J. Underwood, 1984. Underwater microscopic sampling of a sublittoral kelp community. *J. Exp. Mar. Biol. Ecol.,* Vol. *76,* pp. 67-78.

Kenelly, S.J. & A.J. Underwood, 1985. Sampling of small invertebrates on natural hard substrata in a sublittoral kelp forest. *J. Exp. Mar. Biol. Ecol.,* Vol. 89, pp. 55-67.

Lopez-Jumar, E., G. Gonzalez&J. Mejuto, 1986. Temporal changes of community structure and biomass in two subtidal macroinfaunal assemblages in La Coruña Bay, NW Spain. In, *Long-term changes in coastal benthic communities,* edited by C,. Heip, B.F. Keegan & J.R. Lewis, Junk, Dordrecht, pp. 137-150.

Lundalv, T., C.S. Larsson & L. Axelsson, 1986. Long-term trends in algal-dominated rocky subtidal communities on the Swedish west coast: a transitional system? In, Long-term *changes in coastal benthic communities,* edited by C. Heip, B.F. Keegan & J.R. Lewis, Junk, Dordrecht, pp. 81-96.

McDonals, L.L. & W.P. Erickson, 1994. Testing for bioequivalence in field studies: has a disturbed site been adequately reclaimed. In, *Statistics in ecology and environmental monitoring,* edited by D. Fletcher & B.F.J. Manly, University of Otago Press, Dunedin, pp. 183-197.

Peterman, R. M., 1990. Statistical power analysis can improve fisheries research and management. *Can. J. Fish. Aquat. Sci.*, Vol. 47, pp. 2-15.

Ramsey, P. H., 1978. Power differences between pairwise multiple comparisons. *J. Am. Stat. Ass.*, Vol. 73, pp. 479-485.

Ryan, T. A., 1959. Multiple comparisons in psychological research. *Psych. Bull.*, Vol. 56, pp. 26-47.

Scheffé, H., 1959. *The analysis of variance.* Wiley, New York.

Simpson, G. G., A. Roe& R.C. Lewontin, 1960. *Quantitative zoology.* Harcourt Brace, New York.

Snedecor, G.W, & W.G. Cochran, 1989. *Statistical methods, eigth edition.* University of Iowa Press, Ames, Iowa.

Stewart-Oaten, A., W.M. Murdoch & K.R. Parker, 1986. Environmental impact assessment: ‘pseudoreplication’ in time? *Ecology*, Vol. 67, pp. 929-940.

Swihart, R.K. & N.A. Slade, 1985. Testing for independence in animal movements. *Ecology*, Vol. 66, pp. 1176-1184.

Swihart, R.K. & N.A. Slade, 1986. The importance of statistical power when testing for independence in animal movements. *Ecology*, Vol. 67, pp. 255-258.

Tukey, J, W., 1949. Comparing individual means in the analysis of variance. *Biometrics*, Vol. 5, pp. 99-114.

Underwood, A. J., 1981. Techniques of analysis of variance in experimental marine biology and ecology. *Ann. Rev. Oceanogr. Mar. Biol.*, Vol. 19, pp. 513-605.

Underwood, A. J., 1989. The analysis of stress in natural populations. *Biol. J. Linn. Soc.*, Vol. 37, pp.’ 51-78.

Underwood, A. J., 1990. Experiments in ecology and management: their logics, functions and interpretations. *Aust. J. Ecol.*, Vol. 15, pp. 365-389.

Underwood, A. J., 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. *Aust. J. Mar. Freshwat. Res.*, Vol. 42, pp. 569-587.

Underwood, A. J., 1992. Beyond BACI: the detection of environmental impact on populations in the real, but variable, world. *J. Exp. Mar. Biol. Ecol.*, Vol. 161, pp. 145-178.

Underwood, A. J., 1993. The mechanics of spatially replicated sampling programmes to detect environmental impacts in a variable world. *Aust. J. Ecol.*, Vol. 18, pp. 99-116.

Underwood, A. J., 1994a. *Things* environmental scientists (and statisticians) need to know to receive (and give) better statistical advice. In, *Statistics in ecology and environmental monitoring*, edited by D. Fletcher & B.F.J. Manly, University of Otago Press, Dunedin, pp. 33-61.

Underwood, A. J., 1994b. On beyond BACI: sampling designs that might reliably detect environmental disturbances. *Ecol. Appl.*, Vol. 4, pp. 3-15.

Underwood, A.J. & C.H. Peterson, 1988. Towards an ecological framework for investigating pollution. *Mar. Ecol. Progr. Ser.*, Vol. 46, pp. 227-234.

Walters, C.J., 1986. *Adaptive management of renewable resources.* Macmillan, London, 374 pp.

Winer, B. J., D.R. Brown& K.M. Michels, 1991. *Statistical principles in experimental design, third edition.* McGraw-Hill. New York.