Chapter 1
Design and Analysis in Benthic Surveys in Environmental Sampling

Antony J. Underwood and Maura G. Chapman
Centre for Research on Ecological Impacts of Coastal Cities,
University of Sydney, Sydney, Australia

Abstract

Measuring environmental impacts affecting benthic habitats requires detection of specific patterns of statistical interactions in data sampled before and after a potential impact, in the potentially impacted place and in control or reference locations. This is complex because ecological assemblages and populations vary at many spatial and temporal scales. Here, we introduce methods to ensure appropriate, independent replication of sampling at hierarchical scales in space and time. For statistical analysis, the logic of sampling design is critical. Determining precision of estimates and maximising power to detect impacts require care in the design, analysis and interpretation of the relevant data.

Keywords benthic variance, environmental impact, environmental sampling, independence, precautionary principle, precision, replication, scale, sampling design, statistical interaction

1.1 Introduction

Quantified sampling, particularly to test applied and logically structured hypotheses about patterns and processes in marine habitats, is of increasing importance in underpinning the understanding of natural processes and to predict changes in response to environmental influences. There have been numerous advances in soul-searching (Peters, 1991), methods of analysis (Clarke, 1993; Anderson, 2001), commentaries on logic (Underwood, 1990; Resetarits & Bernardo, 1998; Lawton, 1999) and the need for better understanding of environmental impacts (Schmitt & Osenberg, 1996; Sparks, 2000). As a result, it is not possible in a general summary to review comprehensively even the new material, let alone everything relevant to

Methods for the Study of Marine Benthos, Fourth Edition. Edited by Anastasios Eleftheriou.

the topic of improved benthic sampling. Suffice it to say that it is also essential to help with management and conservation of diversity, natural resources, systems and functions. Taking care in the acquisition of quantitative information should, therefore, be of paramount importance to all marine biologists and ecologists.

This chapter, therefore, presents a general overview of the issues concerned. It is not, nor could it be, a ‘cookbook’ of procedures that might work. Rather, it is an attempt to consider fundamental issues of replication, in space and time, the nature of variables examined, issues about designing comparative sampling programmes and so forth.

These topics are considered against a general background of the logical structure underpinning sampling methodology. The issue is a simple one – unless the aims and objectives of any study are clearly identified at the outset, the least damaging outcome will be wastage of time, money and resources. The worst outcome would be a complete lack of valid information on which to build understanding, predictive capacity and managerial/conservatory decision-making. Where aims are vague, designs of sampling are usually (if not always) inadequate, data do not match necessary assumptions, analyses are invalid and conclusions suspect.

In contrast, where aims and purposes are logical, coherent and explicit, it is usually possible to design a robust, effective, efficient and satisfactory sampling programme, which will allow aims to be achieved with minimal uncertainty. This seems such common sense that it does not need to be stated – but common sense indicates that the world is flat and that the sun rotates round the earth. Common sense is not enough.

For example, Hurlbert (1984) published a devastating critique of the failure of many published studies to demonstrate a valid basis for reaching conclusions, because the samples analysed were inappropriately or not at all replicated. His study was confined to the *published* studies, in refereed journals, subject to independent scrutiny. The overall situation, taking into account rejected papers, less intensively scrutinised journals and the flood of unreviewed grey literature, was clearly much worse. It is clear, from practical inspection of more recent literature and through reviewing manuscripts and applications for grants, that the situation (though now better) has not improved substantially (Hurlbert, 2004).

The starting point for studies needing quantitative sampling is that objectives should be clear, the variables to be measured should be defined and the sorts of patterns anticipated in the data should be clearly identified (as testable hypotheses). Wherever possible, as much information as is available should have been collected (and understood) about the operational processes operating, their spatial and temporal scales and about the biological interactions in assemblages and responses to environmental variables. Ideally, the constraints of money, time and equipment should also have all been taken into account. In other words, the professional components of scientific work should all be in place. Under these circumstances, it should be possible to design sampling to achieve minimal probabilities of error in analyses.

As a result, the focus in this chapter is on general issues and procedures to provide help and guidance with setting objectives, formulating hypotheses and designing sampling, particularly for measuring ecological impacts. This will serve as an aide-memoire for contemplating issues of logic when dealing with spatial and temporal variability in biological systems. It will also provide an introduction into the broader literature where many advances have been made in methodologies dealing with the problems of biological complexity in the real world.

1.2 Variability in benthic populations

Surveys must always be designed to take into account the fact that benthic animals and plants are extremely patchy in distribution and abundance. Patchiness is caused by processes external to the assemblage, particularly disturbances and recruitment, in addition to processes operating within the existing assemblage. Although anthropogenic disturbances are often very severe (e.g. large-scale trawling or dredging can cause extreme changes in benthic assemblages; Hall & Harding, 1997; Lindegarth *et al*., 2000), natural disturbances are common and can be important contributors to spatial variability of populations. These vary from small-scale disturbances, e.g. being overturned by waves affects the assemblage on a boulder (Sousa, 1979) and potentially the assemblage in the sediment below it, to large-scale processes, such as the erosion of nearshore sediments (Shanks & Wright, 1986) and the destruction of assemblages in response to large storms (Underwood, 1998).

However, the most important contribution to patchiness is probably due to unpredictable and variable patterns of recruitment (Underwood & Denley, 1984). Both settlement itself and post-settlement mortality typically vary at a hierarchy of spatial scales (Caffey, 1985; Gaines & Bertness, 1992). Patterns at larger scales tend to be more predictable because most species are confined to particular habitats within a biogeographic range (Brown & Gibson, 1983). However, at small scales within patches of habitat, there is considerable variability in recruitment (Keough, 1998), caused by local environmental variation (e.g. topographic features of habitat or localised water currents) or by the existing assemblage itself (e.g. gregarious settlement in response to conspecific adults or consumption of larvae by large numbers of sessile species). The numbers of larvae competent to settle that arrive in any site are themselves influenced by a multitude of external processes, many of which act in the water column well away from the site of settlement. The localised processes, both physical and biological, that influence recruitment are interactive, so recruitment is extremely variable in space and time.

In addition, numerous interactions within the assemblages themselves continually alter patterns of abundances and these effects, too, occur at a range of spatial scales. For example, though predation may decrease abundances at the scale of a shore or habitat, within that habitat predation may eliminate species from certain patches, but leave other patches alone. Feeding by eagle rays or shore birds can create extreme small-scale patchiness in abundances of their prey, although these effects are complicated by environmental factors, such as currents and movement of sediment (reviewed by Thrush, 1999). Even in areas with heavy predation, prey may settle in particular microhabitats, where they can grow large enough to escape predation (Dayton, 1971).

Competition, either for space among sessile animals or plants or for food among mobile animals, also contributes greatly to patchiness of assemblages. Therefore, over-growth or dislodgment of one species by another (Keough, 1984) causes very patchy assemblages of sponges, ascidians and other colonial animals in subtidal habitats and of barnacles and various types of algae in intertidal habitats. Although species that are superior competitors for space or food may eliminate inferior species, the relative strength of interspecific competition may be balanced with that of intraspecific competition. This ensures that neither species is eliminated, but that both persist in very variable and patchy numbers.

Processes such as these are better understood for benthic assemblages living on hard surfaces because, first, the patterns are often readily visible and, second, the processes are relatively easily investigated experimentally. They are, however, also important for assemblages in soft sediments, where recruitment may be equally variable (Skilleter, 1992; Whitlatch *et al.*, 1998) and local disturbances, competition and predation alter local abundances, causing very patchy distributions at a hierarchy of spatial and temporal scales (Morrisey *et al.*, 1992a, 1992b; Ysebaert & Herman, 2002; Fig. 1.1). To estimate abundances of benthic animals and plants accurately, measures must be made at the range of spatial scales relevant for the species, assemblage or process under consideration.

1.3 Appropriate replication

Appropriate spatial replication

Whatever the hypothesis being tested and the ultimate use of the data from sampling, spatial replication is a mandatory component of any benthic study. The large variability in numbers and varieties of benthic species from place to place at many spatial scales (see Section 1.2) creates fundamental problems for determining at what scales replication is necessary.

Consider a relatively simple problem of determining the influence of type of sediment on the numbers and types of benthic species. To keep the example very simple, suppose that a particular species of polychaete is generally more abundant where sediments are coarse than where fine sediments form the major proportion. In any ecological study of these worms in some new area, a relevant hypothesis to be tested is, therefore, that abundance will, on average, be greater in an area of coarse sediment than in a corresponding area of fine sediment. Furthermore, suppose that patches of the different types of sediment are about 800 m in diameter. Finally, as is virtually inevitable, imagine that numbers of worms per m^2 of habitat may vary

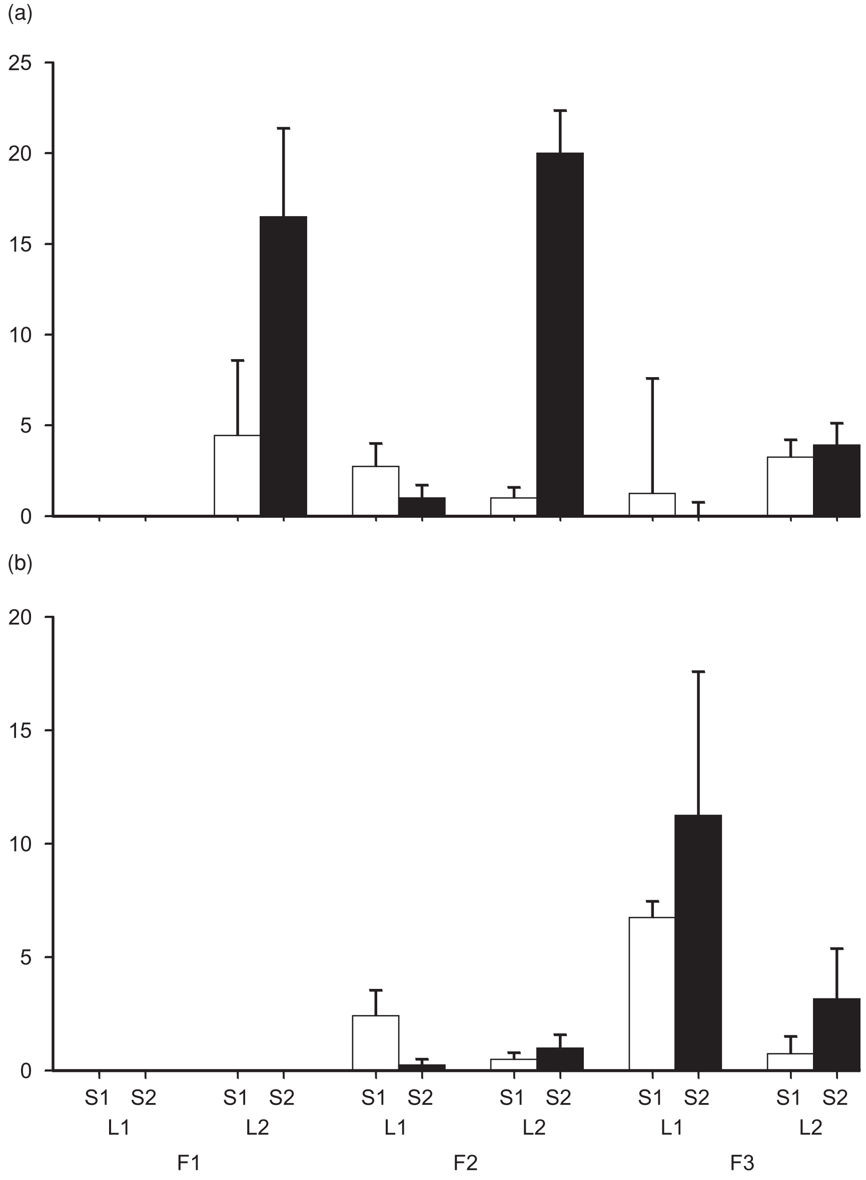


Fig. 1.1 Mean (Standard Error; SE) abundance of two species of amphipods (a) and (b) between sites (S1, S2; tens of metres apart) in each of two locations (L1, L2; 100 m apart) in each of three mangrove forests (F1, F2, F3; kilometres apart). Note that each species shows significant variation at each spatial scale, but these differ between the two species. Patterns of variation at the scale of sites and locations also vary from one mangrove forest to another.

substantially at scales of tens of metres and at scales of hundreds of metres, even in the same type of sediment. Consider a sampling scheme in which 10 box cores are sampled in one patch of coarse sediment and 10 cores are taken from a patch of finer sediment some 2 km away. The mean number of worms is greater in the cores from the coarser sediment, which is consistent with the hypothesis. Nevertheless, this does *not* allow a valid interpretation that the difference is associated with the difference in sediment (as was predicted). Such an interpretation or influence is confounded by the alternative that numbers are different simply because the two areas are 2 km apart. In other words, if samples are taken in two patches of fine or two patches of coarse sediment separated by 2 km, numbers of worms may differ as much as is observed in the two samples shown here. The alternative models explaining why numbers of worms vary are that the sediments differ or that natural variation from place to place, regardless of sediment, caused the numbers to be different. The only way to separate these two models is to predict (from the first model) that differences from one type of sediment to the other are greater than expected from natural variation in either type of sediment.

The study, therefore, requires replication of sites of each type of sediment to estimate natural variation that is not associated with type of sediment. It is sometimes argued that such a study as described is replicated because there were several cores in each patch. This is an example of what Hurlbert (1984) called 'pseudoreplication' – the replicate units in each sample are at the wrong scale and are estimating variability at tens of metres, not at the hundreds of metres at which there are differences among patches of the same type. A better-constructed design would solve the problem by sampling several patches of coarse sediment spaced, say, 2 km apart and several patches of finer sediment at similar spacing, about 2 km from the nearest area of coarse sediment sampled. The two types of patch should be *interspersed*, i.e. chosen to be higgledy-piggledy on a map, so that no systematic trend or gradient makes them different for reasons other than the type of sediment (see Underwood (2000) for examples and illustrations of this issue).

Using such a design, the variation of scales of tens of metres within a patch and at hundreds of metres from patch to patch of the same type of sediment can be estimated. A hierarchical or spatially nested analysis of the data can then be used to test the hypothesis that the difference, on average, between the two types of sediment is greater than the natural spatial variation from patch to patch of the same type. Examples of such analysis for benthic infauna are in Green and Hobson (1970) and Morrisey *et al.* (1992a, 1992b), and the form and structure of analyses are described in detail in Underwood (1997a).

The lengthy description of a very simple case is made necessary by the large number of unreplicated studies, with logically invalid conclusions, that still keep appearing in (or are submitted to and rejected from) ecological journals. It should be noted that better sampling is no substitute for good biological knowledge. So, if the variation in numbers of polychaetes from site to site with finer sediments is known from previous studies to be of the order of tens to hundreds of worms per m^2 and there are 1500 more worms per m^3 in the single coarser site than in the single finer site sampled, it can be argued that this size of difference is much greater than natural variation among sites with finer sediments. In such a case, the conclusion that more worms are found in coarser sediments would be valid.

This argument will, however, only be correct if there has been adequate previous study to demonstrate the general validity (from area to area of the world and from time to time) of the notion that numbers of worms vary naturally by tens to hundreds per metre. It is not usually the case that sufficient information of this type is available. It is, in fact, often the case that previous studies were not replicated at appropriate spatial scales to justify reaching such general conclusions. If such an argument is to be used, it is, therefore, essential to provide the details of the previous studies that might justify it. Often, it is more efficient (and always more valid scientifically, because it does not depend on inductive inferences) to use appropriate replication in any new study.

One further example of problems of logical conclusions from sampling at only one or two spatial scales will be considered. Suppose that dredging is being done at several places in an estuary, to keep channels open for shipping. The sediments in shipping channels are contaminated by heavy metals. Sampling is required in order to detect any impact on benthic infauna in areas around the sites being dredged (not in the sites being dredged – they are not just impacted, but the habitat is actually removed). The concern is that fine sediments with associated heavy metals may wash from the dredged sites to areas up to 100 m away. A replicated study can be designed (as in Fig. 1.2), with several dredged areas and several controls being sampled. In each area, replicated patches of sediment are sampled. Natural spatial variation from patch to patch and area to area is estimated and analysis can reveal any systematic difference between sites near dredging and control sites.

Suppose, however, that the movement of contaminated sediment accidentally released from a dredged area is actually much larger than anticipated. Contaminated fines may now be dispersed over the whole estuary, thus deleteriously affecting all of the control sites, in addition to the sites immediately adjacent to dredging. Changes in benthic fauna due to heavy metals will now occur over all the sites sampled, but there will be no apparent impact because the control sites and the sites next to dredging will not show any differences, when the data are analysed. It would seem that, during the course of the study, there has been an estuary-wide change in fauna not associated specifically with dredging. This sort of situation requires sampling at much larger scales, best of all in other estuaries where dredging is not occurring.

Designs of this type and procedures for analysing the data to detect impacts in such situations were discussed in detail by Green (1979) and Underwood (1992, 1994). The moral of this example is clear – when in doubt about the relevant spatial scale, use a design that can detect changes or differences at one or more of several of the possible scales.

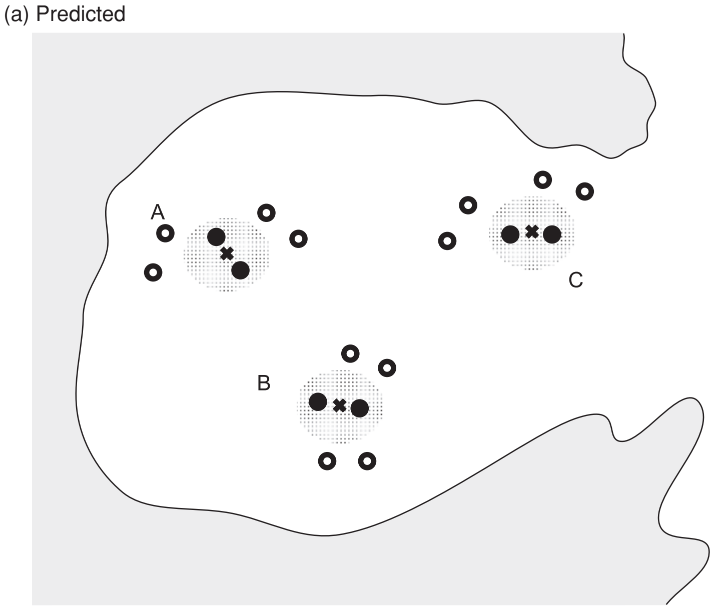


Fig. 1.2 Sampling to detect an impact due to escape of fine sediments from dredging at × in areas A, B and C in an estuary. In each area, there are sampling sites (•) within the distance predicted that sediment would disperse if accidentally released (i.e. over the stippled areas in (a)). Control sites (○) are outside the predicted area of impact and at similar distances apart as the potentially impacted sites. In (b) is the actual, much larger (stippled) area impacted; no impact would be detected because all the controls are also affected.

Appropriate temporal replication

Many hypotheses concern temporal changes, e.g. seasonal patterns of variation or potential changes in populations due to disturbances. A major problem associated with many such studies is the confounding of temporal and spatial variability (Underwood, 1993; Stewart-Oaten & Bence, 2001). For example, to measure seasonal patterns of abundance, a common procedure is to collect a sample, using a set number of replicates, each month (or each season) in one or more places. Tests of temporal patterns then typically compare the abundances from one month (or season) to another to the variability among replicates in the different seasons (using procedures such as analyses of variance). The problem with such comparisons is that variation among seasons is indeed a measure of temporal variation, but variation within a sample is calculated from measures of spatial variation, because the replicates taken at any particular time are all taken at the one time, even though spatially scattered. In such a design, seasonal (or other temporal) patterns are not contrasted against temporal variation within each season, but against spatial variation.

To test for seasonal variation (or other *a priori* selected scales of temporal variation), temporal variation among the factors of interest must be compared to temporal variation *within* each factor of interest. In other words, temporal variation among seasons must be compared to the magnitudes of variation that occur in each season. To measure such variability, it is essential to collect samples several times within each season. With two or more scales of temporal sampling, seasonal or other long-term trends can be identified against background noise. Where there is no measure of shorter-term temporal variation and such variation is large, quite spurious seasonal (or other temporal) patterns will be seen in the data (Fig. 1.3).

Different scales of temporal sampling are extremely important for identifying environmental impacts. Disturbances to the environment may either be short-lived (pulse disturbances) or persist for long periods of time (press disturbances) (Bender *et al*., 1984). The responses of organisms to either type of disturbance may be relatively short-term (i.e. a pulse response); for example, abundances may rapidly increase, but soon drop to normal levels, irrespective of whether the disturbance persists or ceases. Alternatively, populations may show long-term responses (i.e. press responses) to continuing disturbances (because the disturbance continues to exert an effect) or to pulse disturbances (because the disturbance, although it ended long ago, caused a long-term change to another environmental or biological variable). The experimental designs needed to distinguish among pulse and press responses to pulse or press disturbances have been thoroughly described by Underwood (1991) and Glasby and Underwood (1996). All require a sampling design that can measure temporal variation at different temporal scales, measured at the spatial scales relevant to the pulse and press responses.

For examining most environmental impacts and many other ecological hypotheses, the temporal scales of change are not known and can seldom be predicted. The

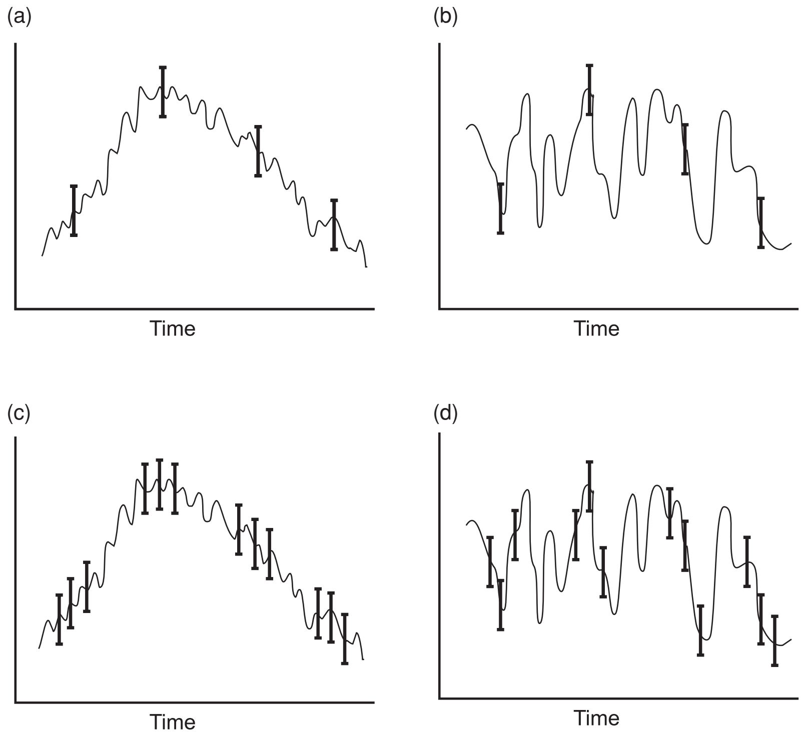


Fig. 1.3 With only a single sample at a series of time intervals a long time apart (e.g. each season), an apparent seasonal pattern could be identified in the variable being measured whether (a) there is, indeed, a long-term seasonal trend or (b) there is considerable short-term variability, but no long-term trend. Short-term temporal sampling is needed within each season. This provides the correct form of within-season replication to measure seasonal changes and identifies (c) long-term trends from (d) background 'noise'.

only way to identify important temporal change is to use a hierarchical temporal sampling scheme. Such designs (Underwood, 1991) not only identify whether any change indicates a pulse or press response but also quantify temporal variances at a number of scales. Impacts that change variances may be more common than and as important as those that change means (Underwood, 1991). If an impact causes much greater fluctuations in abundance (e.g. by altering water currents so that recruitment is more unpredictable), the average abundance of a species may not change over long periods of time; nonetheless the species may be very rare or very common at different times, depending on recent recruitment (Fig. 1.4). If a subsequent impact occurs during a period when abundance is small, such populations may be far more vulnerable to cumulative, but relatively minor impacts, which can then cause local extinctions. Sampling designs that measure changes in mean abundance at different temporal scales, in addition to changes in temporal variances of abundance, are most likely to be successful at identifying environmental impacts and natural ecological variation.

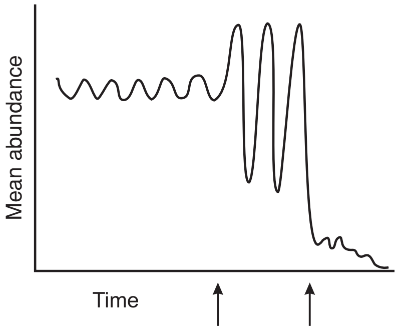


Fig. 1.4 Environmental impacts (at the times indicated by arrows) that affect the variability in abundances of a species, rather than the mean abundance, can cause increased vulnerability to further disturbance. If a subsequent impact occurs during a phase when abundances are very small, the species may easily be driven to local extinction.

1.4 Size of sampling unit

Because abundances typically vary at a range of spatial scales, the size of the sampling unit selected to sample populations is very important in identifying patterns of abundance. For example, consider a population of polychaetes or other small benthic animals that typically aggregate in clusters, about 10 cm in diameter, with the clusters spaced about 20–30 cm apart (Fig. 1.5a). Therefore, some patches of sediment have very large numbers of animals, whereas other patches have very few or no animals at all. Sampling with very large cores or quadrats, for example 50 cm × 50 cm, will give the impression that the animals are very regularly spaced throughout the site. Each quadrat will probably sample one cluster, with perhaps a few individuals from adjacent clusters, giving very similar measures of abundance in each replicate (Fig. 1.5b). This is due to the fact that this size of sampling unit

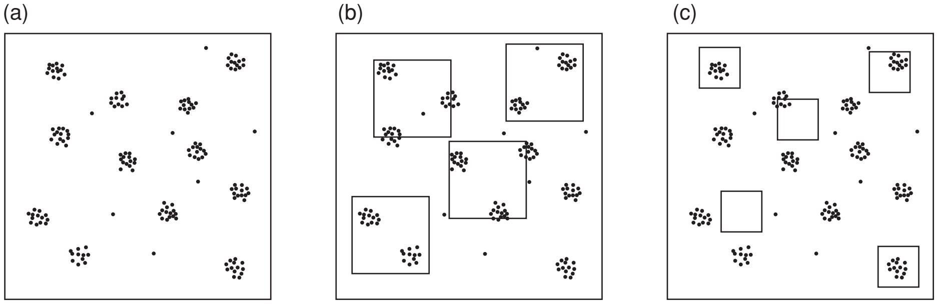


Fig. 1.5 (a) Small benthic animals frequently aggregate into clusters separated from one another. (b) Sampling with units that are much larger than the clusters tends to produce data that suggest a very regular distribution because each unit samples a similar number of individuals. (c) Sampling with smaller units shows the very clustered spatial pattern, with some units sampling clusters and others sampling the bare space among the clusters.

is too large to measure the spatial pattern because the ecological processes causing these patterns are operating at scales of 30 cm or smaller.

Sampling with a quadrat smaller than 30 cm, e.g. 15 cm (Fig. 1.5c), provides a more accurate picture of the spatial variability of these organisms. Some of these units will land near or around clusters, which will thus give very large numbers of the animals. Others will land in spaces among clusters, which will thus give very small numbers of the animals. The average of the replicates should be the same as when large quadrats are used (when scaled to the same area), as long as each set of quadrats representatively samples the population. The variance will, however, be much larger because it will more accurately represent the true pattern of dispersion, thus identifying the scale at which the ecological processes causing these patterns are operating.

In many cases, particularly for organisms in sediments or for very small or cryptic organisms, one cannot observe the patterns of dispersion prior to sampling, so it is not always easy to determine the appropriate size of a sampling unit. In these cases, a pilot study in one site at the start of the experiment, in which a range of sizes of sampling units are used, may identify such patterns and save subsequent, time-consuming efforts. Alternatively, good knowledge of the natural history of the species under investigation, or of other, similar species, or a critical evaluation of the sizes of sampling units others have used, with the reasons why and the patterns identified, may help.

When patterns of abundance of several species occupying the same habitat are being quantified, different-sized sampling units can be used for the different species simultaneously. Therefore, for intertidal and subtidal rocky shores in Australia, a 50 cm × 50 cm quadrat has been shown to provide independent samples that accurately quantify patterns of dispersion and abundance for several of the larger gastropods and barnacles (Underwood, 1981). There are many smaller species that show most variability at the scale of a few centimetres, e.g. littorinid snails (Underwood, 1981). To count these in a 50 cm × 50 cm quadrat is not only time-consuming (because there may be many thousands in a quadrat of this size), but would also not measure variation at the scale that is most important for these small animals. It is more useful, therefore, to count these and similar small species in replicated, smaller subquadrats, scattered representatively over the area of the large quadrat. Similarly, there may be large animals, such as sea urchins or large whelks that may be seldom sampled in a 50 cm × 50 cm quadrat, unless a very large number of quadrats is used. Therefore, these animals may be sampled in large units, e.g. 3 m × 3 m, within which one may sample some 50 cm × 50 cm quadrats, within each of which some 5 cm × 5 cm subquadrats (Fig. 1.6) may be sampled. Each of these scales is relevant to the species being measured and no single scale is appropriate for all species.

Another problem that can arise from the use of inappropriate or a single size of sampling unit may occur when attempting to measure patterns of association between abundances of two species. Consider two species of amphipods living

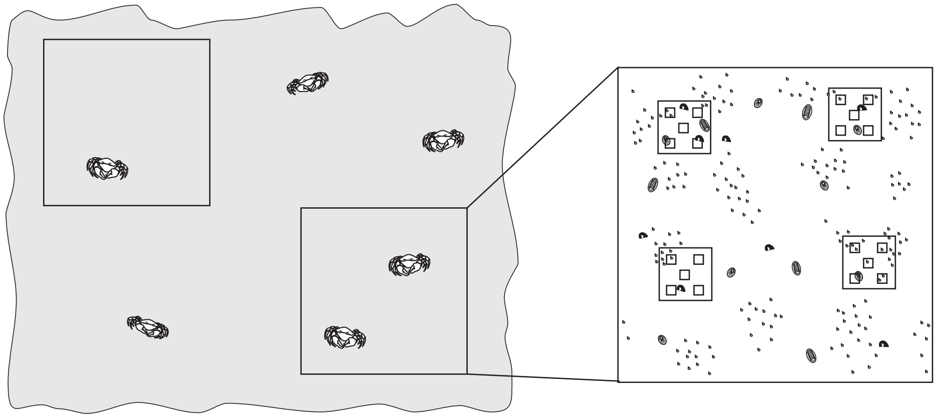


Fig. 1.6 When sampling a suite of species of different sizes and abundances, the most time-effective design is to sample large animals in large quadrats (or other sampling units) scattered over the site, smaller animals in a number of subquadrats, nested inside each of the large quadrats and very small or numerous animals in a set of even smaller areas scattered inside each subquadrat. Such sampling measures variation of the small animals over the same extent of the site as the large animals and allows the variation of the small animals to be partitioned among the different spatial scales.

in shallow, subtidal sediments. If sampled with large cores, say 25 cm in diameter, abundances of the two species may be very strongly positively correlated (Figs. 1.7a, b), i.e. they increase in abundance together. If sampled with small cores, say 10 cm diameter, abundances may be very strongly negatively correlated (Figs. 1.7c, d), i.e. where there are many of Species A, there are few of Species B and vice versa. These two different patterns could be caused because the two species are responding similarly to certain environmental variables, which cause their numbers to increase or to decrease similarly. For example, water depth may have a large-scale gradient across the site and both species are more common in shallower than in deeper water. At the same time, the two species may respond differently to other environmental variables, so that where numbers of one species are large, those of the other species are small and vice versa. For example, there may be small ripples in the sediment across the study site, with one species living up-current and the other down-current of each ripple. This would separate the animals at a scale smaller than 25 cm, thus revealing the negative correlation picked up with the smaller sampling unit (Fig. 1.7e).

Use of only one of the units would lead to the conclusion that the two species were either positively *or* negatively correlated. In fact, they are negatively *and* positively correlated. The different-sized units also direct attention to the scale of the important ecological processes. Therefore, it is necessary to consider factors that cause positive correlation at the scale of the entire site, perhaps over tens of metres, in addition to factors that cause negative correlation at the scale of cm within the site. The opposite pattern could equally be found. For example, species may be positively correlated at a small spatial scale, if they use the same microhabitat. At a large

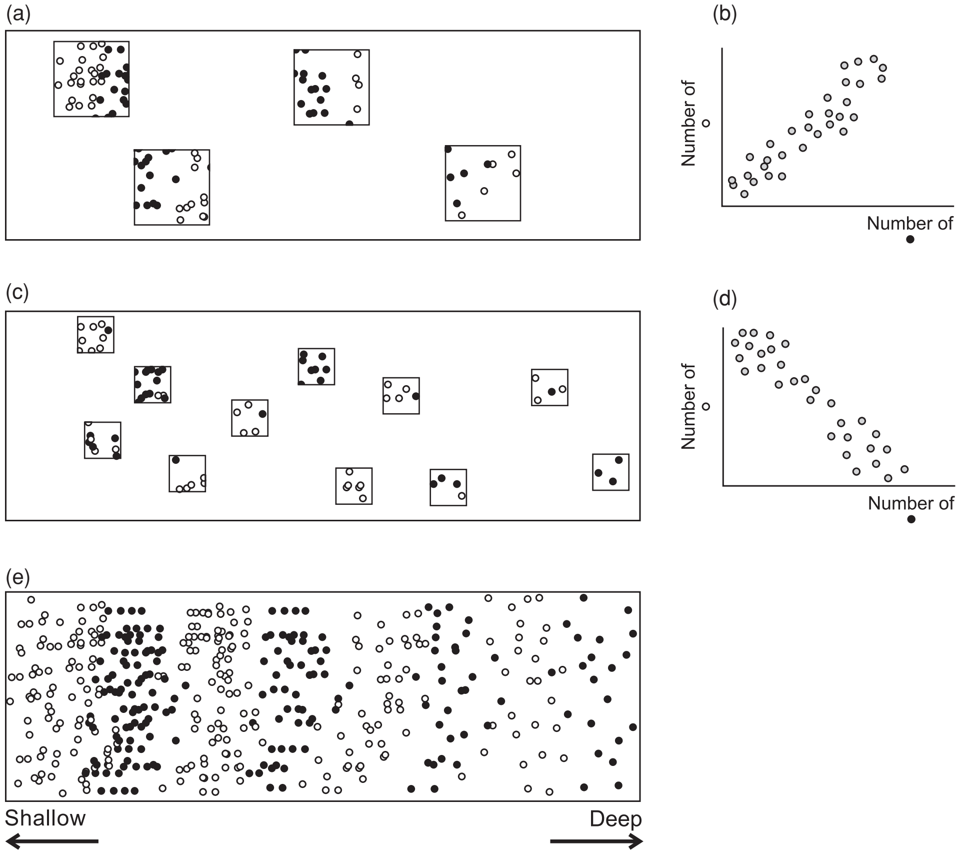


Fig. 1.7 Sampling two species using a large sampling unit (a) shows a positive relationship between the species (b), whereas a smaller unit (c) shows a negative relationship (d). (e) This occurs because the two species are responding to small-scale environmental variables that tend to separate them, while each of them increases in the same way across a broad environmental gradient. Note that the actual pattern in (e) is not visible or known when sampling most types of benthic organisms, so correlations at one scale may not be a good, nor complete, description of relationships between species.

spatial scale, e.g. across a depth gradient, they may be negatively correlated, with one species more common in shallow water and one more common in deeper water.

It is often very difficult to identify individuals of many sessile animals and plants. Therefore, they are usually sampled *in situ* by estimating the percentage of space that they occupy. This can be done by photographing patches of habitat and determining cover from the photographs using a grid of dots or planimetry, or directly in the field. Cover can be estimated from a grid of points, with the species under each point being scored (Underwood, 1981), or by dividing a quadrat into a series of smaller grids and estimating cover in each of these on a 0–4 scale (Benedetti-Cecchi *et al*., 1996). Comparing different methods of sampling cover of sessile organisms in the field, Foster *et al*. (1991) and Benedetti-Cecchi *et al*. (1996) have shown that estimates of cover vary according to the method used and a pilot study may be needed to determine which method gives greater accuracy and precision.

It is clear that the size of the sampling unit used will have a major influence on the spatial patterns that are identified and, hence, understanding of ecological processes. In their extensive review on this topic, Andrew and Mapstone (1987) strongly emphasised the need to consider carefully the sizes (and numbers) of sampling units used in a study to optimise any sampling design, rather than automatically using methods that are reported in the literature, without any idea of their suitability.

1.5 Independence in sampling

Many analytical procedures require that data be independently measured from place to place, time to time and replicate to replicate. This is a particularly important assumption for many univariate procedures (regressions, analysis of variance, chi-squared tests). The issues are not so stark for some multivariate analyses and for many tests where data are permuted to generate distributions of test statistics (Clarke, 1993). Nevertheless, even where the consequences of non-independence may not be clearly understood, they should not be ignored.

Correlations in space occur among data where, for example, density in one area influences what happens in surrounding areas. For example, for the numbers of the two species illustrated earlier (see Section 1.4), there would be positive correlations between data points taken on the peaks of ripples (where numbers of one species are relatively large). There would be negative correlation between any pair of sample units where one is in a trough (where numbers are small) and one is on a peak (where numbers are large).

Many biological processes cause correlations in spatial data. For example, if some species are scattered at random across an area of habitat, there will be no pattern of correlation between the numbers in pairs of sample units, wherever they are placed. If, in contrast, numbers are non-random because predatory fish eat the animals and the activity of fish is concentrated in a few sites, perhaps where there are rocks, there are now patches relatively devoid of the prey animals. Two or more sampling units (cores, grabs, quadrats) landing in one of the areas where predators are active will have small numbers of prey and there will thus be some positive correlation in numbers found in sample units where numbers are small.

Non-independence through time (more commonly called serial correlation) can be a more serious problem. Analysing differences in numbers of animals (numbers of species, or whatever other variable) is fraught with difficulties because the numbers in any given area tend to be correlated with future numbers. Even after trends due to seasonal cycles or mortality through time have been identified in analyses, there is often a tendency for the data at one time to be related to those at the next time, or to those at subsequent times of sampling.

There are two general issues about serial correlation. The first is to determine whether it is present in any given set of data and the second is to determine what can be done about it. A general test for serial correlation is the Durbin–Watson test (Durbin & Watson, 1951), which uses residuals from whatever trend or time-course has been fitted to the data. If e_i represents the residual at time i and there are t times of sampling,

$$d = \frac{\sum_{i=1}^{t} (e_i - e_{i-1})^2}{\sum_{i=1}^{t} e_i^2}$$

is compared with tabulated values when there is no serial correlation. If positive correlation is present, differences between adjacent times of sampling are smaller than expected by chance and d will be smaller than in uncorrelated series of data.

If serial correlation is present and many times of sampling (or distances for spatial analyses) are available, time-series analytical procedures will help. Procedures are described in detail in Box and Jenkins (1976), Diggle (1990) and Cliff and Ord (1973). One possibility is to repeat the Durbin–Watson test at different temporal (or spatial) lags. Thus, compare e_i with e_{i-2}, e_{i-3}, etc., to identify an interval at which there is no longer any noticeable correlation. Data from such distances or times apart would then be sufficiently independent and could be analysed by traditional procedures. This must, however, result in a loss of data and a considerable waste of effort collecting data that are subsequently not useable.

What this means in practice for sampling biota is that, usually, it is not possible to have data for several temporal or spatial samples. Therefore, it is crucial to think in advance about the biological and environmental processes operating that will cause positive or negative correlations in the data. Then, it is often possible to ensure that sampling is done at large enough distances or times to ensure that the correlations do not turn up in the sampled data. Procedures such as hierarchical sampling schemes will also help to allow relevant analysis and some forms of fractal analysis may offer alternative approaches (Leduc *et al*., 1994).

1.6 Multivariate measures of assemblages

Examination of many types of models about natural ecological processes, such as predation or competition, or about changes in response to environmental disturbance or management, requires measures of assemblages of species. Differences in assemblages from place to place, or changes through time, are complex because several variables can change simultaneously. These include the species or taxa found in each sample, their relative abundances and their distribution among replicates within each sample, i.e. their spatial variation. Each of these is equally important in understanding natural ecological processes or changes in response to disturbances. Therefore, analytical tools that can identify changes in assemblages need to take all of these factors into consideration.

Several procedures attempt to reduce such multivariate data into univariate measures that summarise aspects of the entire data set, e.g. species richness, evenness, dominance, etc. None of these measures, however, deals simultaneously with the entire variability of the data. For example, showing that the number of species does not vary from place to place allows no ecological interpretation if it is not known that the identities of the species are different in different places. More useful procedures examine components of a multivariate set of data simultaneously, thus measuring the magnitude of differences between samples in composition of species, abundances and spatial pattern. These procedures are well described elsewhere (Clarke, 1993) and will be introduced only briefly here.

These procedures generally work on the same principle (illustrated in Fig. 1.8). Abundances (biomasses or equivalent measures) are measured in a suite of taxa (Fig. 1.8a) in each of a set of samples, each with a number of replicates. A similarity (or dissimilarity) matrix is then calculated for all pairs of replicates across all of the samples in the matrix (Fig. 1.8b). A number of measures of similarity can be used, but the Bray–Curtis coefficient of similarity,

$$S_{jk} = 100 \left[1 - \frac{\sum_{i=1}^{p} \left| y_{ij} - y_{ik} \right|}{\sum_{i=1}^{p} \left(y_{ij} + y_{ik} \right)} \right]$$

where p is the number of species and y_{ij} and y_{ik} represent the number of species i in any two sample units (units j and k), is generally considered most suitable for ecological data that tend to have many zero values and where abundances tend to be over-dispersed among replicates (Clarke *et al.*, 2006a). The Bray–Curtis measure is also not affected by species that are absent from both of the replicates being compared. When two replicates are identical, the similarity measure is 100% (dissimilarity 0%) and when they have no species in common, similarity is 0% (and dissimilarity 100%).

These measures of (dis)similarity can then be compared within and among samples, assuming replication in each sample, using analyses such as ANOSIM (Clarke, 1993) or PERMANOVA (Anderson, 2001). These test the null hypothesis that the average magnitude in these measures between samples is not greater than it is within samples. These tests have been extended to consider many complex designs, but extreme care should be exercised in interpreting any significant result because, as described above, differences between samples are affected by the species present, their abundances and their occurrences across replicates. These

(a)

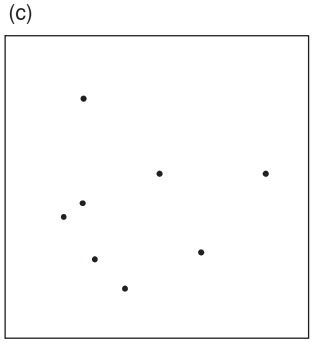
Data

	s1	s2	s3	s4	s5	s6	s7	s8
Species 1	6	12	0	21	0	0	19	11
Species 2	1	0	0	0	2	0	0	0
Species 3	15	3	6	0	0	0	0	0
Species 4	2	1	1	0	1	0	0	0
Species 5	10	20	3	3	3	53	0	0
Species 6	54	55	23	31	79	87	99	56
Species 7	0	12	10	11	3	24	2	0
Species 8	0	13	23	0	0	15	0	0
Species 9	0	1	0	3	0	0	0	3
Species 10	0	5	0	0	0	5	1	0
Species 11	1	0	1	1	0	0	1	0
Species 12	0	0	1	0	3	0	0	0
Species 13	0	0	2	0	0	0	0	0
Species 14	0	0	0	3	0	4	0	0
Species 15	3	6	5	9	0	0	5	0
Species 16	6	0	1	0	0	0	0	0
Species 17	0	0	0	0	0	0	0	3
Species 18	18	1	10	1	3	2	0	3
Species 19	47	34	34	47	64	33	38	44
Species 20	2	1	1	0	0	0	0	0

(b)

Similarity matrix

	s1	s2	s3	s4	s5	s6	s7
s2	68.7						
s3	58.0	66.0					
s4	62.4	67.3	61.4				
s5	67.5	60.2	48.7	59.0			
s6	51.0	71.8	50.0	46.5	63.0		
s7	61.8	66.3	45.5	65.1	73.7	63.4	
s8	75.1	71.8	49.8	72.0	74.1	53.1	73.7



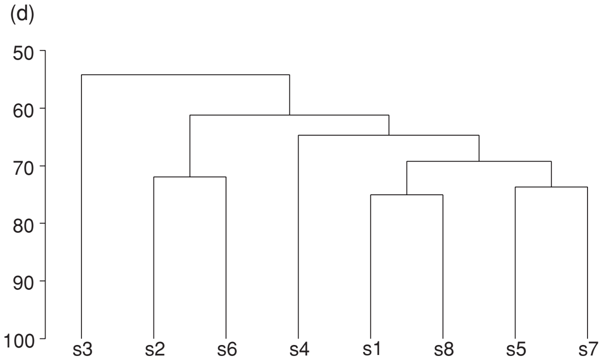


Fig. 1.8 Many useful multivariate analytical procedures convert (a) a matrix of samples, in this case the columns are abundances of a suite of taxa (the rows in the table), or (b) other variables into a matrix of similarities or dissimilarities. This summarises differences across all variables between each pair of samples into a single distance, or measure of difference between the pair. This information can be presented graphically in many ways. (c) Non-metric MultiDimensional Scaling (nMDS) plots illustrate similarity among samples in two- or three-dimensional space, using the ranked distances. Points closer together on the plot are more similar and vice versa. (d) Clustering attempts to identify ‘natural’ groups of samples by identifying which are more similar to each other.

tests do not allow identification of the contribution that each of these types of variation is making to any overall patterns of differences.

Tests of complex designs have other potential statistical problems. For example, differences in dispersion (multivariate variance) within different samples cause difficulties in the interpretation of outcomes of tests for difference among samples. Unlike many univariate procedures, where methods for overcoming such problems have been examined in detail, there has not yet been such development of multivariate procedures.

As with univariate analyses, one must also consider transforming the data prior to analysis. It is common to see the data transformed to $X^{0.5}$ or $X^{0.25}$ in an attempt to minimise the effects of very common species and to allow rare species to contribute more to any patterns of difference. Transformations of different severity (e.g. $X^{0.5}$ is not as severe a transform as $\ln(X)$) affect the patterns found (Olsgard *et al*., 1997) and, thus, the interpretation of any analysis. Another consideration is the level of taxonomic resolution to use. Frequently, very similar results are obtained when organisms are sorted to species, or to genera, or families, or larger groups (Gray *et al*., 1988; Clarke, 1993). This will cause particular difficulties for interpretation when different species in an assemblage are affected differentially by the transformation. For example, species with relatively large abundances are affected more by square root transformation than are species with smaller abundances. Clarke *et al*. (2006b) described procedures to deal with assemblages in which species have different spatial variances among replicate sample units. At the start of any project, spending time to examine the effects of different transformations or levels of taxonomic resolution on the results obtained may well save considerable effort later. It is, however, very important to consider these issues carefully, especially with respect to the model and hypothesis being tested.

Measures of (dis)similarity can also be presented graphically, e.g. in Non-metric MultiDimensional Scaling (nMDS) plots. Such plots attempt to place all the replicates into a two- or three-dimensional diagram, maintaining the relative distances (or measures of dissimilarity) among the replicates (Fig. 1.8c). Replicates that plot closely together contain similar assemblages and replicates that plot further apart contain more dissimilar assemblages. Such plots are particularly useful for illustrating differences among samples and are usually used in conjunction with analyses, such as PERMANOVA.

In addition, samples can be subjected to cluster analyses, of which there are a range of different methods (Clifford & Stephenson, 1975). These attempt to find natural groups of samples, rather than to examine differences between predetermined groups. These groupings are often displayed in a dendrogram (Fig. 1.8d) against a scale of (dis)similarity (e.g. Bray–Curtis similarity), which identifies the degree of similarity that separates samples into different groups.

There are ongoing developments of relevant methods, some of which are very useful in analyses of environmental impacts. A recent summary of relevant techniques is given in Clarke and Gorley (2006).

1.7 Transformations and scales of measurement

There are several reasons why data may need to be transformed prior to analysis. Some parametric and non-parametric analytical procedures (analysis of variance, Kruskall–Wallis tests of ranks) have underlying assumptions, for example, that the data are normally distributed, or that variances among samples are not heterogeneous (i.e. are of the same magnitude). If these assumptions are not met, there is increased possibility of Type I errors (i.e. increased chance of mistakenly finding apparent differences among samples, when they are, in fact, similar; Table 1.4). Although different analyses have different levels of sensitivity to making Type I errors, depending on the particular assumption being violated (Underwood, 1997a), it may be necessary to transform data prior to analysis to meet the assumptions of the analysis (Winer *et al*., 1991).

Three types of ecological processes are responsible for variances being of different sizes among samples. Counts of randomly dispersed animals or plants in quadrats, cores or other sampling units will be distributed approximately as Poisson distributions. This sort of pattern occurs where there are no ecological processes operating to cause the individuals to aggregate or to spread out, so that they are not scattered at random. In a Poisson distribution, the variance equals the mean. Therefore, when comparing samples in which means differ, it is also inevitable that variances will differ (Fig. 1.9a). Comparing samples with very large differences among means is likely to give significant values for heterogeneity of variances (e.g. using Cochran's test). In such situations, where the variance: mean plot is roughly linear, the data can usually be transformed using $\sqrt{X+1}$ to eliminate these problems.

Many ecological processes cause variances to increase exponentially with increases in means so that the data are log-normally distributed. For example, sizes of organisms caused by differing rates of growth, distances moved during dispersal or concentrations of enzymes in animals of different sizes are often log-normally distributed. It is common for most members of a population to grow at relatively similar rates or to disperse relatively similar distances, although a few grow extremely fast or disperse very long distances. These will cause log-normal distributions in the data. Similarly, while small animals may have large concentrations of enzymes, large animals may have small concentrations. If enzyme concentrations are scaled to the size of the animals, measures will be very large for small animals. In cases such as these, the Standard Deviation (SD), not the variance, is approximately linearly related to the mean (Fig. 1.9b). Such data will be more likely to meet assumptions of normality and homogeneity of variances if they are log-transformed prior to analysis (to whichever base seems most appropriate). Unfortunately, if there are zeros in the data, a value (e.g. 0.1 or 1, depending on the range of the data themselves) needs to be added to each datum prior to transformation. Adding these numbers needs careful consideration because this procedure can alter the relative

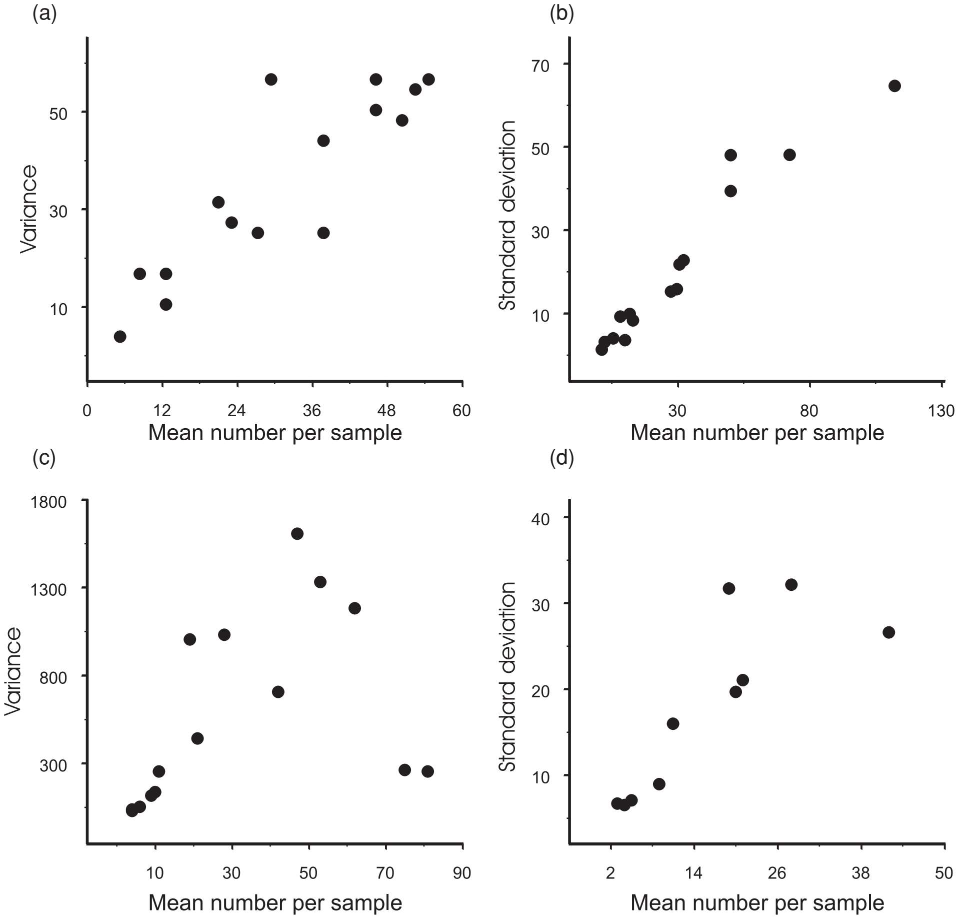


Fig. 1.9 (a) Linear relationship between means and variances in randomly distributed species; (b) linear relationship between Standard Deviations (SDs) and means in log-normally distributed data; (c) relationship between variances and means of proportional (or percentage) data; (d) relationship between SDs and means between constrained sets of proportional (or percentage) data. Parts (a) and (b) are numbers of microgastropods per cobble; (c) and (d) are percentage covers of unoccupied space underneath intertidal boulders.

relationships among means, especially in cases where the data are very variable, e.g. there are many zeros or very small numbers in some samples.

Finally, if data are percentages (or proportions), they may be distributed as binomial distributions. In such cases, variances may be small for samples where means are near the limits of the distribution (e.g. 0% or 100%), but relatively large when samples have means towards the middle of the distribution (e.g. 50%). In such cases, an arc-sine transformation may remove heterogeneity of variances (Fig. 1.9c). When the data do not span most of the range of 0–100%, they will often appear to be distributed like a log-normal distribution, so that a transformation to logarithms may be more appropriate (Fig. 1.9d).

In some cases, even if variances are not heterogeneous and, therefore, transformation is not necessary to meet the assumptions of some statistical test, it may still be appropriate to transform data prior to analysis. Therefore, if the ecological processes influencing the patterns being measured are multiplicative, rather than simply additive, it is appropriate to transform all data to logarithms when testing for differences among mean values. This would ensure that doubling the mean from 10 to 20 from one site to another would give the same measure of difference as doubling from a mean of 100–200. This may be very important, for example, for testing predictions about the effects of pollutants on populations of animals across a number of sites, where the natural density of animals differs greatly from site to site. If the pollutant is hypothesised to decrease densities similarly across a range of natural environmental conditions (and therefore across a range of natural densities), transforming the densities to logarithms will allow for such a test because the test will not be confounded by variation in natural densities. More insights into use of transformations in ecological sampling can be found in Elliott (1977).

In any case, whatever transformation is used for whatever reason, note that transforming data must change the scale over which the data are distributed and, therefore, must alter the relationships between means and variances. Transformations, therefore, also change the meaning or interpretation of any test of a hypothesis. This must be considered very carefully, particularly in multivariate analyses if the same transform is applied to the numbers of a set of species that have very different densities. It is essential that the relationships between the hypothesis being tested, the analytical procedures being used and the scale in which the data are analysed should be clearly maintained to ensure that results of any statistical tests are interpreted logically in terms of ecological processes.

1.8 Data-checking and quality control

Rigorous control of the quality of data is an essential requirement for any research project. This does not only mean collecting relevant, independent, representative data in the field or laboratory for appropriate tests of hypotheses of interest. Very careful translation of the data into the format necessary for analysis is also essential. Nowadays, most data are stored in databases or electronic spreadsheets. It is essential that as much quality control goes into checking that the data in the spreadsheet match those collected in the laboratory or field, as went into collecting the data in the first place. One infallible but time-consuming way to ensure that the data match is to make sure that, after each set of data is entered into the computer, it is also checked by two people, at least one of whom was not involved in the collection of the data. This helps to eliminate many errors that may arise from

Table 1.1 Calculations of Standard Deviations (SDs) from counts of small snails in 15 samples. Note the very large SD for Sample 5 in Column 2, due to a value of 233 being entered instead of 23. When this was corrected (Column 3), Sample 5 still had a large SD, but this was correct. It was due to naturally large abundances and patchiness in this sample.

Sample	SD – not checked	SD – checked
1	0.00	0.00
2	0.45	0.45
3	4.92	4.92
4	0.00	0.00
5	101.04	15.96
6	8.26	8.26
7	7.46	7.46
8	7.60	7.60
9	2.59	2.59
10	7.14	7.14
11	5.45	5.45
12	5.02	5.02
13	4.04	4.04
14	3.42	3.42
15	3.65	3.65

illegible handwriting, fatigue or the sheer boredom associated with entering large amounts of data into databases or spreadsheets.

If it is not possible to develop a foolproof procedure, there are possible short-cuts. For example, if one has a large set of univariate data, such as the sizes of urchins, the diversity of amphipods or the numbers of polychaetes, from a number of sites, it is often useful simply to calculate the SD for each site. If data have been omitted, or very aberrant values have been incorrectly entered, they will cause a very small or a very large measure of SD in a site, relative to other sites (Table 1.1). Therefore, a very unusual SD may indicate an error either in collection of the data in the first place, or an error in transcribing data into the computer. This procedure is quick and easy to use as a form of data-checking, but it will not pick up relatively small errors, only those that cause substantially different estimates of SDs in samples.

When data from a number of species (or variables) have been collected, another form of quality control is to calculate distance measures among the different replicates (e.g. Bray–Curtis dissimilarities for species, or Euclidean distances for abiotic variables). If these are plotted as an nMDS (Non-parametric Multi Dimensional Scaling) plot in two- or three-dimensions, the plot will quickly show outliers, i.e. replicates that have abnormal values for one or more species/variables (Fig. 1.10a). Available procedures to identify which species/variables contribute most to measures of dissimilarity (Clarke, 1993) may help to focus on the particular data that should be checked.

Of course, not all samples with excessively large or small SDs or which form outliers in nMDS plots are necessarily the result of an error. They may simply reflect large natural variability. Similar procedures can also be used to determine whether the sampling strategy is appropriate or not. For example, if replicates

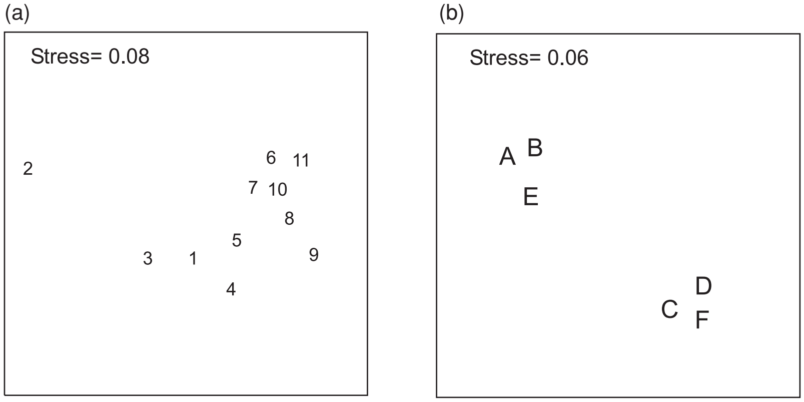


Fig. 1.10 (a) Replicate two in this sample of macrobenthos from a mangrove forest forms an outlier in the Non-metric MultiDimensional Scaling (nMDS) plot because the data for 3 of the 37 taxa in the set of data were omitted from this sample (and were, therefore, erroneously given a value of 0). When this was corrected, the replicates grouped into a single cluster. (b) If the points from a number of replicates in a single sample cluster into two or more distinct groups, it often suggests that the replicates are coming from two distinct strata and the sampling design should be stratified.

cluster into two distinct groups (Fig. 1.10b), it suggests that one might be dealing with two habitats or sets of environmental conditions, each with a different set of taxa. A stratified sampling strategy would probably be better in such a case, with, for example, A, B and E treated as replicates within one stratum and C, D and F as replicates in another (Fig. 1.10b; see Chapman & Underwood, 1999). Similarly, if a species were very aggregated in some sites, but more randomly distributed in others, the SDs would tend to fall into two distinct groups. This too would suggest that sampling should not be done across all sites as if they were equal. Instead, sites should be stratified into two groups (see Section 1.11).

Visual examination of patterns of means, variances, correlation and multivariate data is as important as the statistical tools used to analyse the data. It can provide important information about errors in data and about patterns of variability within and among samples. It should, therefore, be considered as an important precursor to any analysis, rather than looking at the data after statistical differences have been identified.

1.9 Detecting environmental impacts as statistical interactions

Much benthic sampling is done to detect or measure the size of environmental impacts. It is, therefore, worth considering the sorts of sampling designs that are best able to do this. Of course, many situations exist in which optimal designs cannot be done because of lack of time before a disturbance causing an impact, or lack of resources sufficient to achieve adequate spatial replication. Wherever a sampling design cannot be optimal, designing the sampling as carefully as possible will ensure the greatest chance of reliable and unambiguous detection of impacts. By understanding what is needed, greater appreciation can be achieved concerning what is not possible and the consequences for interpretation of data.

The main features of sampling designs necessary to detect impacts have been described in detail by Green (1979) and Underwood (1994, 2000). In an ideal design, there should be data from before to after a disturbance that might cause impacts, so that, if an impact does take place, it can be demonstrated that it appeared after the disturbance purported to have caused it. There must be proper temporal replication before and after the disturbance to provide reliable estimates of average conditions (Bernstein & Zalinski, 1983; Stewart-Oaten *et al*., 1986) and to estimate temporal variance, which might itself be altered by an environmental disturbance (Underwood, 1991).

There must be replicated, undisturbed controls to demonstrate that an impact, if it occurs, is associated with the disturbed area and is not a general phenomenon happening in that habitat, which is not due to that disturbance (Green, 1979). 'Undisturbed' in this context means subject to any other influence or process except the particular disturbance under investigation (Underwood, 2000). Control sites should be replicated to prevent confounding of interpretations (Underwood, 1992; see the section entitled 'Appropriate spatial replication').

Finally, as convincingly demonstrated by Green (1979), an environmental impact should always be considered as an ecological interaction. An impact is a change from before to after a disturbance (planned or accidental), which is not the same in the disturbed area as in an undisturbed control area. Therefore, there is a lack of consistency in the temporal pattern of change of the variables being measured in the disturbed area and the patterns of change in the controls (Fig. 1.11). Accordingly, in order to analyse impacts, it is necessary to design sampling that will provide data that can be analysed to detect and interpret statistical interactions.

One type of ideal design is illustrated in Fig. 1.11 and the analysis is shown in Tables 1.2 and 1.3. The situation concerns the construction of sewage outfalls along a coastline. They are to be situated on rocky headlands with fast currents, deep water and coarse offshore sediments. Two outfalls were examined, with two corresponding control areas with similar features of habitat (rocky headlands with fast currents, deep water and coarse sediments). Any coast-wide change in benthic fauna that is not due to discharge of sewage will affect the control areas and the outfall locations. An impact will cause a different type of change where there are outfalls from where there are no outfalls (hence an interaction in the spatial difference between outfall and control locations from before to after the outfalls are commissioned).

Data are collected three times (essentially chosen at random) over two years before the outfalls begin to discharge sewage and then three times over a period of two years, starting two years after the outfalls are commissioned (Fig. 1.11). At

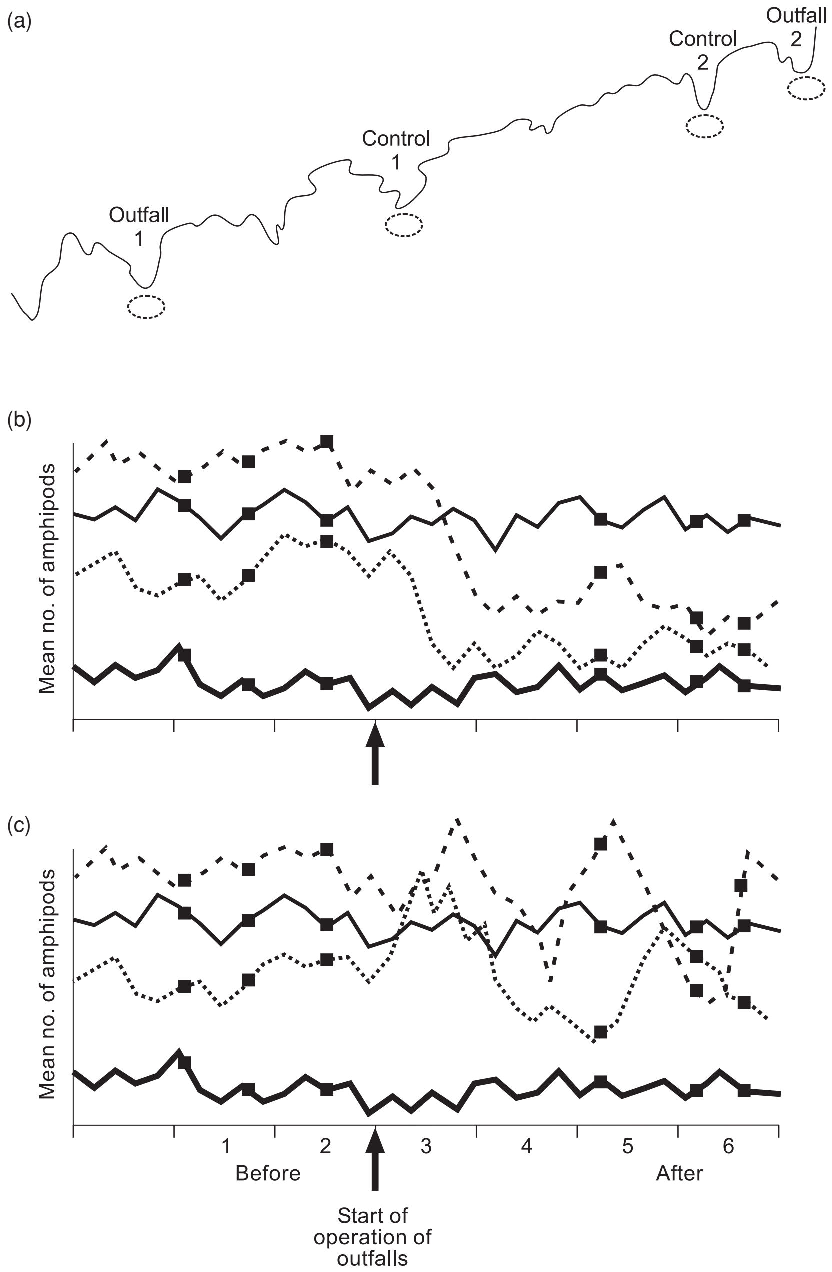


Fig. 1.11 (a) Sampling to detect impacts from construction of coastal sewage outfalls. Two outfalls (—, . . .) and two similar control sites (—, **—**) are sampled 3 times (■) before and again after the outfalls begin to discharge. (b) Data that would indicate a large-scale, consistent 'press' impact. (c) Data that would identify shorter-term fluctuating impacts.

Table 1.2 Framework for analysis of hypothetical sampling to detect impacts due to construction of sewage outfalls (illustrated in 1.11). *O* indicates the comparison of mean numbers of an amphipod between outfall and control locations and is a fixed factor (see Winer *et al*., 1991; Underwood, 1997a). There are two (randomly chosen) locations of each type, indicated by *L*(*O*). Sampling before and after the outfalls were built represents a fixed factor; there are 3 randomly chosen times of sampling before and again after the outfalls start operation. $n = 5$ replicate cores are sampled at each time in each location.

Source of variation		Degrees of freedom	Test
Outfalls versus controls	$= O$	1	Irrelevant
Before versus after	$= B$	1	Irrelevant
$O \times B$		1	Indicates large-scale, long-term impact
Locations (outfall or control)	$= L(O)$	2	Irrelevant
$B \times L(O)$		2	Indicates different impacts at the two outfalls
Times of sampling (before or after)	$= T(B)$	4	Irrelevant
$O \times T(B)$		4	Indicates possible fluctuating impacts
$L(O) \times T(B)$		8	Indicates possible fluctuating impacts that differ between outfalls
Residual		96	
Total		119	

Table 1.3 Framework for an asymmetrical (beyond Before-After/Control-Impact (BACI)) analysis of sampling to detect impacts due to construction of a single sewage outfall. There is now no replication of outfall locations, but all other details are as in Table 1.2.

Source of variation		Degrees of freedom	Test
Outfalls versus controls	$= O$	1	Irrelevant
Before versus after	$= B$	1	Irrelevant
$O \times B$		1	Indicates large-scale, long-term impact
Locations (control)	$= L(C)$	1	Irrelevant
$B \times L(C)$		2	Indicates different impacts at the two outfalls
Times of sampling (before or after)	$= T(B)$	4	Irrelevant
$O \times T(B)$		4	Indicates possible fluctuating impacts
$L(C) \times T(B)$		4	Indicates possible fluctuating impacts that differ between outfalls
Residual		72	
Total		89	

each time of sampling, a certain number, say five, replicate areas of sediment would be taken in each location and the animals of some species of interest (perhaps an amphipod) collected from each area.

From such data, the analysis in Table 1.2 can be completed. A large and consistent press impact (Bender *et al.*, 1984; Underwood, 1989, 1991) would cause an interaction identified as $O \times B$ in Table 1.2. A shorter-term, more fluctuating response (perhaps a pulse response; Bender *et al.*, 1984) would appear as an interaction $O \times T(B)$ in Table 1.2. Where an impact has a different effect on fauna in the two locations with outfalls, there will be an interaction $B \times L(O)$ or $L(O) \times T(B)$ depending on the rates of change caused by the sewage. Impacts causing changes in temporal variance rather than changes in mean abundance of animals (Fig. 1.11b) can also be analysed (Underwood, 1991, 1994).

Of course, in most cases, there is only one planned disturbance (in the present example, only one outfall is planned). Consequently, there is a loss of replication of the disturbed ('outfall') location and correspondingly less certainty about any impact. There must, however, still be replication of control locations, ensuring that extreme forms of confounding do not happen. Such asymmetrical designs are still analysable and interpretable (Table 1.3) and have been called 'beyond BACI' designs (Underwood, 1992) to distinguish them from earlier, spatially unreplicated and confounded BACI designs (Stewart-Oaten *et al.*, 1986). The acronym BACI is for 'Before-After/Control-Impact' sampling, where there is one (and only one) disturbed (called 'Impact') location and only one control area. The disadvantages and logical problems with BACI designs are well known (Underwood, 1992; Smith *et al.*, 1993).

In many situations, there can be no sampling before a disturbance, either because it is accidental (such as an oil spill) or because the impacts were unforeseeable and start to appear only after the disturbance has started. Such situations can still be analysed (Chapman *et al.*, 1995; Glasby & Underwood, 1996), but, again, there may be uncertainty about the cause of the impact. Without previous data from before the event, there is less certainty that the impact started after the disturbance.

There is now a substantial body of reviews of literature on sampling and analysis of environmental impacts, starting from Green (1979) and, more recently, Spellerberg (1991), the papers in Schmitt and Osenberg (1996) and Sparks (2000). There are still many problems in analyses of complex samples for multivariate measures, but interactions caused by impacts can, in the simplest cases, be analysed by permutation procedures (Anderson, 2001).

1.10 Precautionary principles and errors in interpretations

Precaution has become a guiding goal for decision-making in a framework of ecological sustainability (see Cameron & Abouchard, 1991; Wynne, 1992). Despite different views by lawyers, managers and scientists on what precautionary

Table 1.4 Two types of error in conclusions from statistical tests.

	As a result of sampling, the null hypothesis is:	
	Rejected	Retained
	You believe there has been an environmental impact	You believe there has been no environmental impact
Unknown to you, the null hypothesis is:		
TRUE	TYPE I ERROR With probability α	CORRECT DECISION
FALSE	CORRECT DECISION	TYPE II ERROR With probability β

principles mean (Dovers & Handmer, 1995), there is nevertheless consensus that precaution can be defined in terms of environmental decisions. Given uncertainty about the scientific information underpinning managerial decisions (and, although often forgotten, the uncertainty in all the architectural, economic, engineering, legal, political and social information), it is important to remember that mistakes do occur in the interpretation of analyses of quantitative ecological data. In statistical terms, there are two major types of mistake (Table 1.4). Type I errors occur when a null hypothesis is rejected by a statistical procedure, but which is, in fact, true. Type II errors occur when a null hypothesis is false (i.e. some alternative is true), but is retained. In environmental decision-making, the hypothesis put forward is that there will be an impact due to some disturbance. For other types of ecological studies, hypotheses include predictions such as the following:

(i) Based on the observation that faunal assemblages vary with depth and composition of sediments varies with depth, it can be proposed that differences in composition of sediments cause the difference in assemblages. Therefore, a relevant hypothesis is that, for any particular composition of sediments, similar fauna will be found at all depths. Also, at any depth, the fauna in a particular type of sediment will be similar to that in the same type of sediment at other depths.

(ii) Given previous observations of fewer amphipods in areas where rays feed, the model can be proposed that predation or disturbance by the rays decreases the number of amphipods. This leads to the hypothesis that areas where rays are experimentally prevented from entering will develop larger numbers of amphipods than in corresponding control areas.

Sampling is then done to test the hypothesis (hypotheses) and data are analysed statistically. A statistical procedure includes choosing a probability (α) of Type I error – the probability that the data collected cause rejection of the null hypothesis as the outcome of the test, even though the hypothesis is, in fact, correct. So, an impact could apparently be found where there is no impact; different fauna seem to be present in similar sediments, whereas in reality, there is no difference. The presence of rays seems to make a difference, but in fact, the numbers of amphipods are not influenced by rays.

Type II errors exhibit the opposite situation – no impact is found, even though one has occurred. A Type II error occurs when fauna do actually differ, but no difference is found in the samples; or if the sampled numbers of amphipods are not significantly different in the presence and absence of rays, even though rays are, in fact, important predators. The probability of making Type II errors (β) cannot be assessed in most studies.

In environmental studies, failure to detect an impact (a Type II error) is much more serious than erroneously claiming to find one (Type I) where there is none. Where an impact is believed to have occurred, more work will usually be done (to check, to determine its extent, to test further hypotheses about its consequences). In such a case the mistake will almost certainly be detected. In contrast, where a real impact has not been detected, no particular follow-up will occur and environmental degradation will continue. Precautionary principles, therefore, are based on the assumption that errors should be of Type I and not of Type II (Gray, 1990, 1996; Peterman & M'Gonigle, 1992; Underwood, 1997b).

There is not enough space here to demonstrate how to reduce the probability of Type II errors. Details can be found in Green (1989), Peterman and M'Gonigle (1992) and Underwood (1997b). The probability of not making a Type II error, i.e. of correctly rejecting a null hypothesis (finding an impact) when it is false (there *is* an impact), is $(1 - \beta)$ and is called the power of a test (see Table 1.4). Power is increased where:

(i) α, the probability of Type I error is increased;
(ii) the intrinsic variance, or 'noise', in the measurements is small;
(iii) the size of the effect is large, i.e. large differences are expected if the null hypothesis is false;
(iv) the sizes of samples are large.

Of these, the variance is a property of the system being measured (but see Section 1.11). α and the sizes of samples (numbers of replicate sites, times of sampling, sample units in each site) are chosen by the experimenter. Thus, they should be chosen to achieve large power, i.e. great capacity to reject null hypotheses (to find differences where they exist).

The key to power analysis is the effect-size. To calculate the power of any sampling and analysis requires that the amount of difference among times, places, habitats, experimental treatments, etc., be specified in advance. Thus, the hypothesis that rays influence amphipods is devoid of real information. What is required is a statement, based on available knowledge, that, if rays decrease the numbers of amphipods, removing them from some experimental areas will cause an increase of *X* amphipods per unit area of habitat. Where the original observations provoking the study were that fewer amphipods are present where rays feed (see above), *X* is known from these observations (it is the observed difference hypothetically caused by rays).

For the assessment of environmental impacts, effect-sizes are more difficult to establish. A good yardstick is that the effect-sizes (how much difference in the densities of the affected species) or how much change in any other relevant variable is expected (if there really is an impact) should be based on what responses would be triggered by the discovery of that size of impact. So, suppose that an impact causes a reduction of 10% in numbers of crustaceans in some mud flat near an industrial outfall and constitutes an impact, but there is no proposed change in regulation or management. In contrast, suppose that a reduction of 60% would trigger immediate regulatory responses. It is, therefore, very important to be able to detect a 60% difference, but of limited or no value to be able to find much smaller impacts.

In this case, a sensible effect-size would be somewhere around 50% change and the power of the study should be made large (say $>$90%) in order to have a good chance of finding the impact if it occurs (Green, 1989; Underwood, 1997a). Where resources (money, time, equipment (i.e. money, money, money!)) are insufficient to allow adequate replication to achieve large power, α (the probability of Type I error) should be increased in order to achieve reductions in β (the probability of Type II error). Mapstone (1995) has described in detail how to trade off the lack of resources (increasing β) by increasing the chance of mistakenly finding impacts (increasing α) in the assessment of environmental impacts.

1.11 Precision and the size of samples

As indicated earlier (see Section 1.10), the power of tests is a function of the sizes of samples used. In general, the precision of an estimate from a certain sample is increased as a function of the size of sample. A typical measure of precision is the standard error, which is (sample variance/size of sample)$^{1/2}$ and clearly decreases (so precision increases) as size of sample increases. Wherever possible (and it is always desirable), a maximal acceptable imprecision should be specified. It is possible to estimate the variance of the variable being measured and thereby to calculate how many replicates should be included in a sample to achieve the necessary precision.

There are, however, several features of design of sampling that help to increase precision of estimates of abundances of organisms. The first is stratification. Wherever it is possible to make a ‘map’ of abundances (from previous studies in the literature or from pilot studies) stratification of sampling will often substantially reduce imprecision. As a simple example, let us suppose that it is generally known that a particular species of sea urchin is generally more abundant (per box core) in areas of very coarse sediment than where sediments are finer. Let us suppose that in the study area, about 25% of the seafloor is composed of coarse sediments, in several large patches. The remaining areas are finer sediments. There are sufficient

Table 1.5 Sampling urchins in an estuary with two types of sediments. In (a), $n = 16$ cores are taken at random over the whole area; some are in coarse, some are in finer sediments. In (b), $n = 4$ cores are taken in the areas of coarse sediment and $n = 12$ in areas of finer sediment. Stratification substantially increases the precision of the estimate of mean number per core.

(a) Numbers of urchins in 16 cores

- 19, 100, 77, 13, 1, 15, 20, 17, 90, 77, 8, 22, 8, 78, 14, 29
- Mean = 36.8; variance = 1156; SE = 8.5

(b) Numbers of urchins

- Coarse sediment: $n = 4$
  - 102, 80, 100, 95
  - Mean = 15.2; variance = 289; SE = 8.5
- Finer sediment: $n = 12$
  - 8, 10, 5, 6, 26, 23, 25, 26, 17, 8, 1, 27
  - Mean = 15.2; variance = 81.1; SE = 2.8
- Combined sample:
  - Mean = 35.6; SE = 2.6

funds to take a total of 16 cores to estimate the average numbers of urchins per m^2 of seafloor, as part of an ongoing study of disturbances due to trawling. Suppose you now take the samples at random positions across the area, resulting in the data in Table 1.5a. This gives an estimate of mean abundance of 36.8, with Standard Error (SE) = 8.5.

If, instead, the sampling were to be stratified – i.e. you took 25% (or 4 of the 16 samples) in the areas of coarse sediment and the remainder in the other habitat, then you would have the data in Table 1.5b. Now, the mean number of urchins is estimated separately in each habitat, with separate estimates of imprecision, which are then combined for the whole area to give a mean abundance of 34.9, with SE = 2.5. This latter method is much more precise because it removes the variation among replicate cores that is due to the systematic differences between the two types of habitat. Of course, it also gives explicit information about the densities of urchins in each of the two habitats, which may or may not be useful (depending on the actual hypotheses being tested).

Combining the samples from each habitat, as done here, is only valid if the variances in numbers of urchins in the two areas are measured independently. Alternative methods, in particular how to stratify samples so that the number of replicates taken in each stratum is proportional to the variance in the stratum, have been described in detail (Cochran & Cox, 1957; Cox, 1958) and other issues in marine ecology were reviewed by Andrew and Mapstone (1987).

A different method of expressing imprecision is to calculate the Confidence Interval (CI) for a particular mean estimated from a sample. A 95% CI indicates a range of values that have a 95% probability of including the true mean being estimated. If the CI is small, the unknown mean has been estimated with a fair degree of precision.

Table 1.6 Sampling an amphipod in three habitats to demonstrate precision of combined samples. Data (hypothetical) are numbers in $n = 5$ cores in each habitat. Variances were homogeneous (Levene's test, F with 2, 12 df = 1.18, $P > 0.30$).

		Habitat		
		1	2	3
Replicate	1	10	28	12
	2	14	31	9
	3	3	33	3
	4	5	26	0
	5	2	19	1
Mean		6.8	27.4	5.0
Variance		25.7	293	31.2
SE		2.3	2.4	2.3
95% CI (4 df)		6.3	6.7	6.5
Combined variance		28.7	28.7	28.7
Combined SE		2.4	2.4	2.4
95% CI (12 df)		5.2	5.2	5.2

A 95% CI for a sample mean $\bar{X}$ from a sample of size n is

$$\bar{X} \pm t_{0.05,(n-1)\mathrm{df}} \times \sqrt{\text{Sample variance}/n}.$$

As size of sample (n) increases, the CI will decrease because the SE (i.e. (variance/mean)$^{1/2}$) is smaller and because there are more degrees of freedom (df) (i.e. ($n - 1$) increases), giving a smaller value of t (from Student's t-distribution, in this case choosing $\alpha = 0.05$, the probability of Type I error).

In any study of differences in means through time (seasons, before/after events, etc.) or across space (habitats, patches, depths, etc.), it makes sense to combine samples, so that the estimates of variance around each mean come from a combined estimate of variance, with more df than from each sample alone. Thus, putting the samples together as in an analysis of variance improves the precision of sampled estimates of means, as illustrated in Table 1.6. For this to be valid the variances of the various samples should be similar and this can be tested by various procedures. The example in Table 1.6 considers samples of $n = 5$ cores from each of three habitats. The combined CI is much smaller than the individual ones. To achieve the same precision (i.e. the same small CI) for each habitat separately would have required samples of $n = 7$ for each habitat, thus increasing the required work 1.4 times.

1.12 Gradients and hierarchies in sampling

There has been debate about appropriate sampling designs to use for analyses of influences along an environmental or other gradient. For example, when testing a hypothesis about the influence of discharge of fresh water from an estuary, it is appropriate to sample along a gradient from the mouth of the estuary. Similarly and more obviously, sampling down a depth gradient requires the sites to be along the gradient.

In other cases, the situation is not so clear-cut and whether or not sampling along a gradient is the best option depends entirely on the hypothesis being tested. Furthermore, even if a gradient is to be sampled, the spacing of samples is dependent on the precise issue under examination.

To provide an illustration of the issues, consider a simple case of discharges of contaminants from an outfall on the coast (see Underwood, 2000, for more details). In all three situations considered here, it is proposed that pollution due to the contaminants will be revealed by differences in the assemblages of animals in sediments near to, as opposed to far away from, the outfall.

In the first case, the design of the outfall included modelling of the probable dispersal of contaminants and their dilution away from the pipe (Fig. 1.12a). The requirement of the ecological sampling is to test the hypothesis that concentrations of contaminants in the animals in sediments do, indeed, conform to the modelled pattern of distribution. In this case, samples must be taken along the gradient, at fairly regular intervals, to have the greatest chance of detecting departures from the prediction.

In the second case (Fig. 1.12b), the issue of concern is not the actual spatial pattern of contamination, but the distance from the outfall at which concentrations of contaminants fall below a certain critical value. For example, if the outfall is discharging pesticides in run-off from housing, there may be, for human safety, a maximal allowable concentration of contaminants in scallops on the seafloor. It is, therefore, important to focus attention on the limits within which it is unsafe to harvest scallops for human consumption. In this case, sampling effort should be entirely focused on the areas where modelling indicates that concentrations of contaminants are below the critical threshold. The resources and effort should be used to identify the 'safe' boundary rather than the actual spatial gradient of contamination.

Mostly, however, there are major disconnections between gradients of contamination and actual biological responses, i.e. pollution (Phillips, 1978; Spellerberg, 1991; Keough & Black, 1996; Raimondi & Reed, 1996). Such disconnections are due to the inertia of biological systems, i.e. animals or assemblages not being affected by that contaminant, so that there is no response despite chemical signals. Alternatively, biological responses can occur when concentrations of contaminants are minimal, or undetectable, because of bioaccumulation in individual animals. Finally, populations may be unaffected by pollution due to widespread dispersal of their larvae, maintaining populations from elsewhere (Underwood & Peterson, 1988). In such cases, analyses along gradients are not likely to be the best approach, because there is no clear indication of the course or extent of the gradient. An effective alternative is to sample at sites near the outfall to make comparisons with distant sites that are chosen to be control sites because they are located far enough away to be unaffected by the outfall (Fig. 1.13a). As discussed above (see

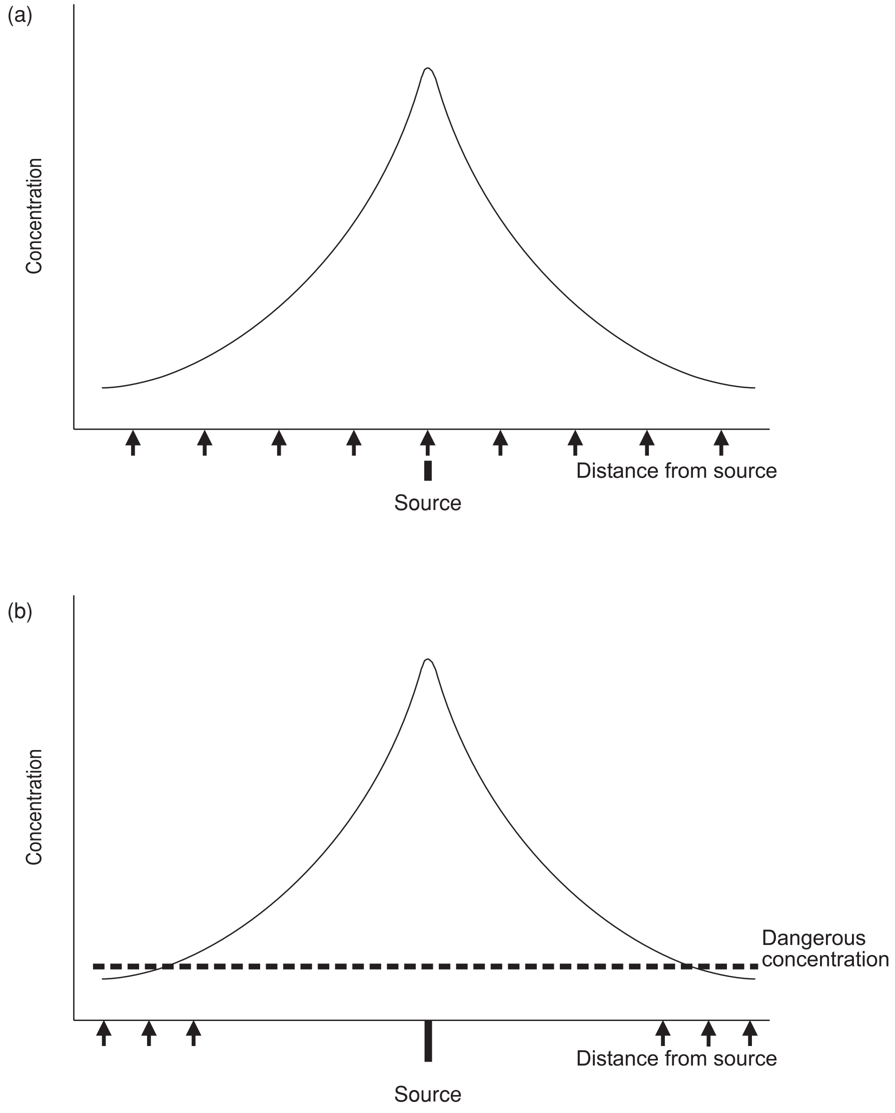


Fig. 1.12 Sampling relevant to gradients of some influence (e.g. pollution from an outfall pipe). (a) The issue is the spatial pattern of decreasing impact based on predicted spatial dilution of contaminants, so sampling along the gradient away from the outfall is necessary. (b) The major concern is risk to health due to pollution being over a defined ‘safe’ limit in animals at some distance from the outfall. Sampling is, therefore, concentrated around the areas where this safe limit is supposed to occur.

the section entitled ‘Appropriate spatial replication’), uncertainty about the scale of the potential pollution requires sampling at more than one spatial scale.

Another design that may be useful concerns tests of hypotheses about differences in assemblages resulting from some installation or intrusion into a habitat. For example, it may be proposed that fauna in sediments near to rocky reefs are different

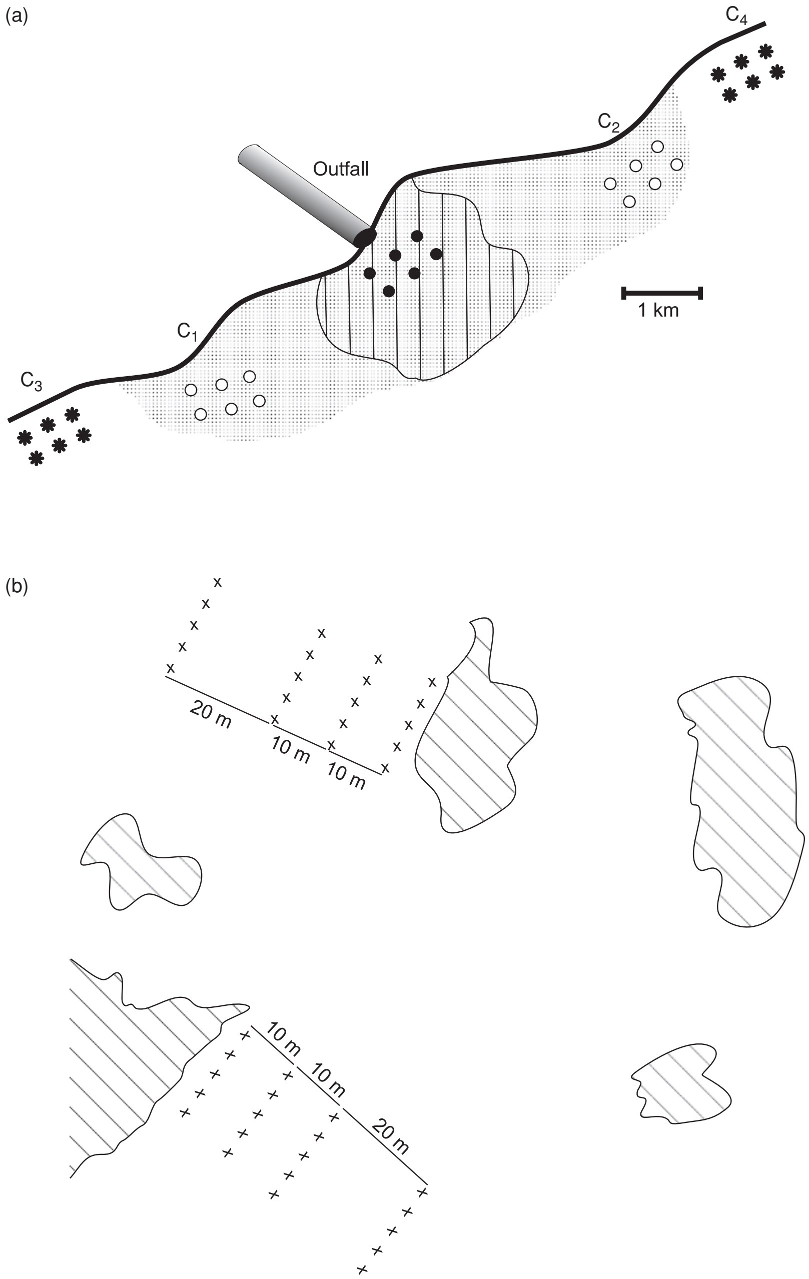


Fig. 1.13 (a) Sampling to handle uncertainty about the scale of an impact due to an outfall. It has been proposed that any impact would only extend over a small distance (stippled), so samples are taken from that area and from controls on either side, outside the area expected to be affected. The area influenced may, however, be much larger (shaded), so other control samples are taken at much greater distances from the outfall. (b) Sampling to test the hypothesis that assemblages in sediments are influenced by the presence of a rocky reef (reefs are shaded). For two replicate reefs, $n = 5$ cores are taken next to the reef (0 m) and at 10, 20 and 40 m from it. The reef will cause samples at 0–10 to differ more than 10–20 or 0–20 to differ more than 20–40 (i.e. more than similar distances where there are no reefs; for more details, see the text).

from those further away. The reasons for these differences may be physical (waves, water flow, sediments may all be affected by the presence of a reef) or biological (predatory fish are associated with reefs and therefore eat animals close to a reef, but not those further away). In this case, sampling is needed to test the hypothesis that assemblages close to a reef are different from those further away. Samples should then be taken near a reef and at some distance away, say, 10 m (Fig. 1.13b).

The hypothesis is that any difference is greater than normal spatial differences, so a further set of samples is taken 20 m in from the reef. If the reef influences the fauna, the difference in uni- or multivariate measures of a species or the assemblage from the samples at the reef and at 10 m will not be the same as the natural variation at that scale, i.e. the difference from 10 to 20 m. If the scale of potential influences is not known in advance, other distances can be sampled. Thus, a further sample at 40 m allows comparison of the differences between samples at the reef and at 20 m with the natural variation between 20 and 40 m (Fig. 1.13b). A detailed example of the use of this design was described by Kelaher *et al*. (1998) for infauna in sediments near boardwalks in mangrove forests.

The point to be understood is that sampling designs must be modified for each particular situation and the general principle to guide designs must be to have a flexible framework (Green, 1993), responsive to the needs of the particular hypotheses being tested (Underwood, 1997a).

1.13 Combining results from different places or times

Sometimes, a particular study may be quite small so that, taken on its own, it cannot reveal much about a process under investigation. It is, therefore, important to think about mechanisms for combining the outcomes of such studies in a meta-analysis, i.e. a combination of results of similar tests of the same hypothesis in different places and times and conditions. General issues about ecological meta-analyses were pioneered by Gurevitch *et al*. (1992).

Here, some procedures are introduced. First, in many experimental studies, the samples are quite small and the outcomes are not clear, because the tests are not powerful enough to provide unambiguous interpretations. There are two general procedures available for combining the results of several small tests. Let us suppose that the outcomes of experimental treatments (say, mean numbers of some polychaete at five different depths) are available from several small studies (perhaps six different cruises have collected such data, but each only had two or three samples from each depth). The means from the five depths can be ranked from smallest to largest for each study and summed. This gives the frequency out of six studies of each depth having the smallest, second, third, fourth or the largest number of worms. Such data can be analysed by a test devised by Anderson (1959) to determine non-random patterns in the rankings (see Underwood & Chapman, 1992, for a detailed example).

Second, there may be a test of a certain hypothesis in several studies. Each gives a probability of the data being likely to be due to chance errors in sampling, but only a few (or none) are significant. The probabilities from these small tests can be combined using Fisher's (1935) test

$$C = -2\sum_{i=1}^{k} \log_e (P_i)$$

where k is the number of tests and P_i is the probability from each test ($i = 1, \ldots, k$). C is distributed as χ^2 with $2k$ df if there are, in fact, no differences in the set of tests. Thus, large values of χ^2 would cause the rejection of the null hypotheses over a set of tests. An example is given in Table 1.7.

Finally, in some situations, several small tests have been done, but the meta-analysis requires them to be weighted, for example, to reflect different amounts of habitat across a study area. Suppose, for example, it is proposed that numbers of species of crustaceans are greater in some estuary in areas with seagrasses than in patches of open, sandy sediments among the seagrass beds. Three species of seagrass form beds in the estuary and one of them (*Posidonia*) occurs in densely or sparsely covered beds. Thus, there are four habitats of interest. In the estuary, there are different areas of each type of habitat.

Samples of $n = 5$ replicate cores were taken among each type of seagrass and $n = 5$ cores were taken in bare patches within the bed. For each such pair of samples, a 1-tailed t-test was done. There were significantly more crustaceans among seagrass in only one of the four tests (Table 1.7), but the probabilities were quite small in several tests. To test the hypothesis of a difference between seagrass and bare areas over the whole estuary, the tests were combined by Fisher's (1935) method, giving $C = 18.29$, with 8 df. This was significant at $P < 0.02$, indicating greater numbers of worms among seagrasses across the whole set of data. This test does not take into account the different areas occupied by the different habitats. So, a Stouffer–Liptak procedure (Folks, 1984) was used, which weights the outcomes

Table 1.7 Example (using hypothetical data) of meta-analysis of numbers of worms per core ($n = 5$ cores per sample) among seagrass and in empty patches in seagrass in an estuary. The proportion of all seagrass made up by each habitat is given and is the weighting for the Stouffer–Liptak procedure (for all other details, see text).

Habitat	Proportion (= weight)	Seagrass	Bare patches	$t_{\text{1-tail, 8 df}}$	P	Z_I
Dense *Posidonia*	0.40	62 (10.1)	51 (9.7)	0.79	0.227	0.749
Sparse *Posidonia*	0.30	15 (2.7)	11 (2.9)	1.01	0.170	0.954
Zostera	0.13	23 (3.0)	18 (3.2)	1.13	0.146	1.054
Halophila	0.17	15 (1.5)	9 (1.9)	2.49	0.019	2.075
$C = 18.28$						
$Z_w = 1.98$						

of tests so that large habitats count for more than do smaller ones in a test of the estuary as a whole. The test statistic is

$$Z_W = \sum_{i=1}^{k} W_i . Z_i \Bigg/ \sqrt{\sum_{i=1}^{k} W_i^2}$$

where Z_i is the standard normal score for P_i, the probability in each test and W_i, is the weighting for the ith habitat. The weighting is the proportional area of the estuary occupied by that habitat, so that widespread, common habitats count for more in the outcome. The probability of getting a value as large or larger than Z_W can be found from the standard normal distribution. In this case, $Z_W = 1.98$, with $P < 0.03$. There is a major difference between areas with seagrass and empty areas across all habitats. A more detailed example from marine ecological data may be found in McDonald *et al*. (1993).

Thus, even where resources do not allow for large amounts of replication and adequate power, it is possible to combine the outcomes of repeated or similar studies, provided they are designed in logically comparable ways and have specific, well-identified hypotheses so that the procedures for combining them are valid.

1.14 Conclusions

It should be clear from the above that there are important and close connections between biological knowledge about systems, logical development of ideas, models and hypotheses, design of sampling, analysis of data and interpretations of the information. Of course, like everything else in science (and life itself), there will be continuing disagreements about the role and purpose of hypothesis-testing in marine research (see Stewart-Oaten, 1996; Suter, 1996). This can be a rich debate or a sterile one. Here, it would be a distraction, so it will receive no mention. Suffice it to say that those who favour risk analysis in environmental assessments and those who favour estimation of magnitudes of differences among sites, times, habitats, conditions, rather than structured tests of formally explicit hypotheses would surely all agree about the need for care in developing sampling programmes. Without advance thought about appropriate scales of replication, it is no more possible to put valid CI around estimates of mean numbers of animals (or any other parameter being estimated) than it is to do a valid statistical test of some hypotheses about a parameter. Estimators and other describers of the world have the same issues for design as do experimentalists testing hypotheses in advancing scientific understanding.

The issues are similar, whatever problem is being investigated. Biological systems are variable in space and time at many scales, due to many interacting processes. It is the responsibility of biologists and ecologists to understand the consequences of such variation and the ensuing interactions and non-independent patterns so that sampling can be planned to take all of the issues into account. Where 'ideal' sampling is not possible, what is uncertain in the data can and must be considered before any conclusions are reached. This will ensure that unplanned difficulties or accidental loss of samples do not prevent interpretations being of value. If major problems are inevitable with some design, alter the design or investigate more preliminary hypotheses in order to unravel the problems.

This introductory assessment of some of the issues is selective and seriously incomplete – there are many other issues. Its purpose is to identify the sorts of issues that should be considered when planning any programme of sampling in marine benthic systems, whatever the issue of concern in the study. Understanding the themes discussed should, at least, provide warnings about pitfalls and some of the vocabulary needed to translate a particular study into meaningful questions to ask statistical advisors.

The problems of pollution, fragmentation and destruction of habitat, overharvesting of resources, restoration of degraded habitats, conservation of biodiversity, control of introduced species, global warming and rises in sea level, etc., are vast and urgent. Never has there been such a need for good scientific understanding and advice about what to do and how, when and where. This science deserves the best scientific practice, so improving logic, design, analysis and interpretation of studies is an urgent and ongoing task for marine scientists. Complacency and unprofessionalism will continue to undermine the role of science and will continue to slow down the implementation of solutions to currently urgent problems. Getting designs of sampling right entails getting problems identified and understood and provides links to valid analysis and interpretation. Improved sampling designs are the key to improved scientific contributions to social needs.

Acknowledgements

Preparation of this paper was supported by funds from the Australian Research Council through the Centre for Research on Ecological Impacts of Coastal Cities. We are grateful to M. Button and V. Mathews for help with preparation of the figures.

References

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

Anderson, R.L. (1959) Use of contingency tables in the analysis of consumer preference studies. *Biometrics*, **15**, 582–590.

Andrew, N.L. & Mapstone, B.D. (1987) Sampling and the description of spatial pattern in marine ecology. *Annual Review of Oceanography and Marine Biology*, **25**, 39–90.

Bender, E.A., Case, T.J. & Gilpin, M.E. (1984) Perturbation experiments in community ecology: theory and practice. *Ecology*, **65**, 1–13.

Benedetti-Cecchi, L., Airoldi, L., Abbiati, M. & Cinelli, F. (1996) Estimating the abundance of benthic invertebrates: a comparison of procedures and variability between observers. *Marine Ecology Progress Series*, **138**, 93–101.

Bernstein, B.B. & Zalinski, J. (1983) An optimum sampling design and power tests for environmental biologists. *Journal of Environmental Management*, **16**, 335–343.

Box, G.E.P. & Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*. Holden Day Inc., San Francisco, CA.

Brown, J.H. & Gibson, A.C. (1983) *Biogeography*. C.V. Mosby Co., St Louis, MO.

Caffey, H.M. (1985) Spatial and temporal variation in settlement and recruitment of intertidal barnacles. *Ecological Monographs*, **55**, 313–332.

Cameron, J. & Abouchard, J. (1991) The precautionary principle: a fundamental principle of law and policy for the protection of the global environment. *Comparative Law Review*, **14**, 1–27.

Chapman, M.G. & Underwood, A.J. (1999) Ecological patterns in multivariate assemblages: information and interpretation of negative values in ANOSIM tests. *Marine Ecology Progress Series*, **180**, 257–265.

Chapman, M.G., Underwood, A.J. & Skilleter, G.A. (1995) Variability at different spatial scales between a subtidal assemblage exposed to the discharge of sewage and two control assemblages. *Journal of Experimental Marine Biology and Ecology*, **189**, 103–122.

Clarke, K.R. (1993) Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, **18**, 117–143.

Clarke, K.R. & Gorley, R.N. (2006) *Primer V6: User Manual*. Primer-E Ltd, Plymouth.

Clarke, K.R., Somerfield, P.J. & Chapman, M.G. (2006a) On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. *Journal of Experimental Marine Biology and Ecology*, **330**, 55–80.

Clarke, K.R., Chapman, M.G., Somerfield, P.J. & Needham, H.R. (2006b) Dispersion-based weighting of species counts in assemblage analyses. *Marine Ecology Progress Series*, **320**, 11–27.

Cliff, A.D. & Ord, J.K. (1973) *Spatial Autocorrelation*. Pion Ltd, London.

Clifford, D.H.T. & Stephenson, W. (1975) *An Introduction to Numerical Classification*. Academic Press, New York.

Cochran, W.G. & Cox, G. (1957) *Experimental Designs*, 2nd edn. John Wiley & Sons, New York.

Cox, G. (1958) *The Planning of Experiments*. John Wiley & Sons, New York.

Dayton, P.K. (1971) Competition, disturbance and community organization: the provision and subsequent utilization of space in a rocky intertidal community. *Ecological Monographs*, **41**, 351–389.

Diggle, P.J. (1990) *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

Dovers, S.R. & Handmer, J.W. (1995) Ignorance, the precautionary principle, and sustainability. *Ambio*, **24**, 92–97.

Durbin, J. & Watson, G.S. (1951) Testing for serial correlation in least-squares regression. *Biometrika*, **38**, 159–178.

Elliott, J.M. (1977) *Methods for the Statistical Analysis of Samples of Benthic Invertebrates*. Freshwater Biological Association Scientific Publication No. 25, Reading.

Fisher, R.A. (1935) *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Folks, J.R. (1984) Combination of independent tests. In: *Handbook of Statistics 4. Nonparametric Methods* (eds P.R. Krishnaiah & P.K. Sen), North-Holland, New York, pp–72–94.

Foster, M.S., Harrold, C. & Hardin, D.D. (1991) Point versus photo quadrat estimates of the cover of sessile organisms. *Journal of Experimental Marine Biology and Ecology*, **146**, 193–204.

Gaines, S.D. & Bertness, M.D. (1992) Dispersal of juveniles and variable recruitment in sessile marine species. *Nature*, **360**, 579–580.

Glasby, T.M. & Underwood, A.J. (1996) Sampling to differentiate between pulse and press perturbations. *Environmental Monitoring and Assessment*, **42**, 241–252.

Gray, J.S. (1990) Statistics and the precautionary principle. *Marine Pollution Bulletin*, **21**, 174–176.

Gray, J.S. (1996) Environmental science and a precautionary approach revisited. *Marine Pollution Bulletin*, **32**, 532–534.

Gray, J.S., Aschan, M., Carr, M.R., Clarke, K.R., Green, R.H., Pearson, T.H., Rosenberg, R. & Warwick, R.M. (1988) Analysis of community attributes of the benthic macrofauna of Frierfjord/Langesundfjord and in a mesocosm experiment. *Marine Ecology Progress Series*, **46**, 151–165.

Green, R.H. (1979) *Sampling Design and Statistical Methods for Environmental Biologists*. John Wiley & Sons, Chichester.

Green, R.H. (1989) Power analysis and practical strategies for environmental monitoring. *Environmental Research*, **50**, 195–205.

Green, R.H. (1993) Application of repeated measures designs in environmental impact and monitoring studies. *Australian Journal of Ecology*, **18**, 81–98.

Green, R.H. & Hobson, K.D. (1970) Spatial and temporal structure in a temperate intertidal community, with special emphasis on *Gemma gemma* (Pelecypoda: Mollusca). *Ecology*, **51**, 999–1011.

Gurevitch, J., Morrow, L.L., Wallace, A. & Walsh, J.S. (1992) A meta-analysis of competition in field experiments. *American Naturalist*, **140**, 539–572.

Hall, S.J. & Harding, M.J. (1997) Physical disturbance and marine benthic communities: the effects of mechanical harvesting of cockles on non-target benthic infauna. *Journal of Applied Ecology*, **34**, 497–517.

Hurlbert, S.H. (2004) On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. *Oikos*, **104**, 591–597.

Hurlbert, S.J. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

Kelaher, B.P., Chapman, M.G. & Underwood, A.J. (1998) Changes in benthic assemblages near boardwalks in temperate urban mangrove forests. *Journal of Experimental Marine Biology and Ecology*, **228**, 291–307.

Keough, M.J. (1984) Dynamics of the epifauna of the bivalve *Pinna bicolor*: interactions among recruitment, predation, and competition. *Ecology*, **65**, 677–688.

Keough, M.J. (1998) Responses of settling invertebrate larvae to the presence of established recruits. *Journal of Experimental Marine Biology and Ecology*, **231**, 1–19.

Keough, M.J. & Black, K.P. (1996) Predicting the scale of marine impacts: understanding planktonic links between populations. In: *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats* (eds R.J. Schmitt & C.W. Osenberg), Academic Press, San Diego, CA, pp 199–234.

Lawton, J.H. (1999). Are there general laws in ecology? *Oikos*, **84**, 177–192.

Leduc, A., Prairie, Y.T. & Bergeron, Y. (1994) Fractal dimension estimates of a fragmented landscape: sources of variability. *Landscape Ecology*, **9**, 279–286.

Lindegarth, M., Valentinsson, D., Hansson, M. & Ulmestrand, M. (2000) Interpreting large-scale experiments on effects of trawling on benthic fauna: an empirical test of the potential effects of spatial confounding in experiments without replicated control and trawled areas. *Journal of Experimental Marine Biology and Ecology*, **245**, 155–169.

Mapstone, B.D. (1995) Scalable decision rules for environmental impact studies: effect size, Type I, and Type II errors. *Ecological Applications*, **5**, 401–410.

McDonald, L.L., Erickson, W.P. & Strickland, M.D. (1993) Survey design, statistical analysis, and basis for statistical inference in coastal habitat injury assessment: Exxon Valdez oil spill. In: *Exxon Valdez Oil Spill: Fate and Effects in Alaskan Waters* (eds P.G. Wells, J.N. Butler & J.S. Hughes), American Society for Testing and Materials, Philadelphia, PA, pp 296–311.

Morrisey, D.J., Howitt, L., Underwood, A.J. & Stark, J.S. (1992a) Spatial variation in soft-sediment benthos. *Marine Ecology Progress Series*, **81**, 197–204.

Morrisey, D.J., Underwood, A.J., Howitt, L. & Stark, J.S. (1992b) Temporal variation in soft-sediment benthos. *Journal of Experimental Marine Biology and Ecology*, **164**, 233–245.

Olsgard, F., Somerfield, P.J. & Carr, M.R. (1997) Relationships between taxonomic resolution and data transformations in analyses of a macrobenthic community along an established pollution gradient. *Marine Ecology Progress Series*, **149**, 173–181.

Peterman, R.M. & M’Gonigle, A.B. (1992) Statistical power analysis and the precautionary principle. *Marine Pollution Bulletin*, **24**, 231–234.

Peters, R.H. (1991) *A Critique for Ecology*. Cambridge University Press, Cambridge.

Phillips, D.J.H. (1978) The use of biological indicator organisms to quantitate organochlorine pollutants in aquatic environments: a review. *Environmental Pollution*, **16**, 167–229.

Raimondi, P.T. & Reed, D.C. (1996) Determining the spatial extent of ecological impacts caused by local anthropogenic disturbances in coastal marine habitats. In: *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats* (eds R.J. Schmitt & C.W. Osenberg), Academic Press, San Diego, CA, pp 179–198.

Resetarits, W.R. & Bernardo, J. (eds) (1998) *Experimental Ecology: Issues and Perspectives*. Oxford University Press, Oxford.

Schmitt, R.J. & Osenberg, C.W. (eds) (1996) *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats*. Academic Press, San Diego, CA.

Shanks, A.L. & Wright, W.G. (1986) Adding teeth to wave action: the destructive effects of wave-borne rocks on intertidal organisms. *Oecologia*, **69**, 420–428.

Skilleter, G.A. (1992) Recruitment of cerithiid gastropods (*Rhinoclavis* spp.) in sediments at One Tree Reef, Great Barrier Reef. *Journal of Experimental Marine Biology and Ecology*, **156**, 1–21.

Smith, E.P., Orvos, D.R. & Cairns, J. (1993) Impact assessment using the before-after-control-impact (BACI) model: concerns and comments. *Canadian Journal of Fisheries and Aquatic Science*, **50**, 627–637.

Sousa, W. (1979) Disturbance in marine intertidal boulder fields: the nonequilibrium maintenance of species diversity. *Ecology*, **60**, 1225–1239.

Sparks, T. (ed.) (2000) *Statistics in Ecotoxicology*. John Wiley & Sons, Chichester.

Spellerberg, I.F. (1991) *Monitoring Ecological Change*. Cambridge University Press, Cambridge.

Stewart-Oaten, A. (1996) Goals in environmental monitoring. In: *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats* (eds R.J. Schmitt & C.W. Osenberg), pp. 17–28. Academic Press, San Diego, CA.

Stewart-Oaten, A. & Bence, J.R. (2001) Temporal and spatial variation in environmental impact assessment. *Ecological Monographs*, **71**, 305–339.

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

Suter, G.W. (1996) Abuse of hypothesis testing statistics in ecological risk assessment. *Human and Ecological Risk Assessment*, **2**, 331–347.

Thrush, S.F. (1999) Complex role of predators structuring soft-sediment macrobenthic communities: implications of changes in spatial scale for experimental studies. *Australian Journal of Ecology*, **24**, 344–354.

Underwood, A.J. (1981) Structure of a rocky intertidal community in New South Wales: patterns of vertical distribution and seasonal changes. *Journal of Experimental Marine Biology and Ecology*, **51**, 57–85.

Underwood, A.J. (1989) The analysis of stress in natural populations. *Biological Journal of the Linnean Society*, **37**, 51–78.

Underwood, A.J. (1990) Experiments in ecology and management: their logics, functions and interpretations. *Australian Journal of Ecology*, **15**, 365–389.

Underwood, A.J. (1991) Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. *Australian Journal of Marine and Freshwater Research*, **42**, 569–587.

Underwood, A.J. (1992) Beyond BACI: the detection of environmental impact on populations in the real, but variable, world. *Journal of Experimental Marine Biology and Ecology*, **161**, 145–178.

Underwood, A.J. (1993) The mechanics of spatially replicated sampling programmes to detect environmental impacts in a variable world. *Australian Journal of Ecology*, **18**, 99–116.

Underwood, A.J. (1994) On beyond BACI: sampling designs that might reliably detect environmental disturbances. *Ecological Applications*, **4**, 3–15.

Underwood, A.J. (1997a) *Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance*. Cambridge University Press, Cambridge.

Underwood, A.J. (1997b) Environmental decision-making and the precautionary principle: what does this mean in environmental sampling practice? *Landscape and Urban Planning*, **37**, 137–146.

Underwood, A.J. (1998) Grazing and disturbance: an experimental analysis of patchiness in recovery from a severe storm by the intertidal alga *Hormosira banksii* on rocky shores in New South Wales. *Journal of Experimental Marine Biology and Ecology*, **231**, 291–306.

Underwood, A.J. (2000) Trying to detect impacts in marine habitats: comparisons with suitable reference areas. In: *Statistics in Ecotoxicology* (ed. T. Sparks), John Wiley & Sons, Chichester, pp 279–308.

Underwood, A.J. & Chapman, M.G. (1992) Experiments on topographic influences on density and dispersion of *Littorina unifasciata* in New South Wales. *Proceedings of the Third International Symposium on Littorinid Biology* (eds J. Grahame, P.J. Mill & D.G. Reid), The Malacological Society of London, London, pp 181–195.

Underwood, A.J. & Denley, E.J. (1984) Paradigms, explanations and generalizations in models for the structure of intertidal communities on rocky shores. In: *Ecological Communities: Conceptual Issues and the Evidence* (eds D.R. Strong, D. Simberloff, L.G. Abele & A. Thistle), Princeton University Press, Princeton, NJ, pp 151–180.

Underwood, A.J. & Peterson, C.H. (1988) Towards an ecological framework for investigating pollution. *Marine Ecology Progress Series*, **46**, 227–234.

Whitlatch, R.B., Lohrer, A.M., Thrush, S.F., Pridmore, R.D., Hewitt, J.E., Cummings, V.J. & Zajac, R.N. (1998) Scale-dependent benthic recolonization dynamics: life stage-based dispersal and demographic consequences. *Hydrobiologia*, **375/376**, 217–226.

Winer, B.J., Brown, D.R. & Michels, K.M. (1991) *Statistical Principles in Experimental Design*, 3rd edn. McGraw-Hill, New York.

Wynne, B. (1992) Uncertainty and environmental learning: reconceiving science and policy in the preventative paradigm. *Global Environmental Change*, **2**, 111–127.

Ysebaert, T. & Herman, P.M.J. (2002) Spatial and temporal variation in benthic macrofauna and relationships with environmental variables in an estuarine, intertidal soft-sediment environment. *Marine Ecology Progress Series*, **244**, 105–124.