Biostatistics Unit, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand.

LEVELS OF EVIDENCE IN STUDIES OF COMPETITION, PREDATION, AND DISEASE

Summary: The primary aim of a scientific investigation is to find the most likely model for a situation out of a host of alternative explanations. The strength of evidence provided by anecdote, logical argument, mathematical modelling, observation, and designed studies (manipulative and observational) are discussed and the effectiveness of randomisation and orthogonal designs in separating hypotheses compared. Pseudoreplication is shown to be often misunderstood. It consists of two concepts: the importance of adequate replication and the independence of the sampling units. While replication is necessary to separate out the effects of different factors and to provide an error term for inference, contrary to popular belief independence of sampling units is not necessary. Finally the interpretation of evidence is discussed and the distinction made between formal and informal generalisation to a population.

Keywords: Levels of evidence; randomisation; orthogonal design; blocked design; pseudoreplication; inference.

Introduction

The primary aim of a scientific investigation is to find the most likely model for a situation out of a host of alternative explanations. We usually start out with many ideas and end up with a few hypotheses but these are usually the ones that are most difficult to distinguish between. Our aim is to thin that down to one with no plausible alternatives.

I am assuming that for this symposium our specific aim is to establish the existence and importance of one of the population processes (predation, competition, or disease) in influencing the numbers of a target species. I will review some of the various levels of evidence that are available to field ecologists faced with this kind of problem. Many sources of information can produce weak evidence at best, so I stress the role of statistical modelling and experimental design for efficiently distinguishing between biologically relevant hypotheses. In this context I discuss the concept of pseudoreplication and try to clear up some misconceptions that surround it.

Terminology

Though this paper has no pretensions to being a philosophical discussion (I want it to be useful) we do have to clarify what I will mean by three words:

Process: The interaction itself, whereby one species eats, infects or competes with another. Though these are the primary subjects for this symposium, mutualistic relationships are more difficult to study than any of these, but may be just as important to the survival of native species (particularly plants - e.g., the role of mycorrhizae and pollinators).

Effect: The interaction is of importance to a population ecologist if the consequences affect the population size. This will usually be by changes in birth, death, immigration, or emigration. As we will consider below, these can interact in complex ways, so even a large change in any one of these may actually have effectively no influence on the size of the study population in the next generation.

Evidence: The information used to separate hypotheses and exclude some as implausible. Weak evidence forces one to say: this evidence is consistent with a number of hypotheses; based on this evidence we have no logical reason for choosing one over another. Good evidence lets us say: only this hypothesis is consistent with the facts; the other hypotheses are contradicted. Of course this only remains true until someone else comes up with a new alternative hypothesis. But that's the nature of science, perceived truth changes: today's dogma, tomorrow's bad joke. For this symposium we can define levels of evidence as the probability of being wrong if we assert the existence and importance of an effect on the basis of the information in front of us.

Evidence therefore allows us to throw the risk of being wrong into the balance with the other factors we need to consider before taking action. Here there is a distinction between scientists as scientists and

scientists as managers. As we shall see, the nature of evidence required to produce scientifically rigorous conclusions is not the same as that required to initiate action in an applied situation, because managers will not be using purely scientific criteria to make their decision. In both science and management we must balance the probability of being wrong against the risks of acting on the estimation of the effect, but risk assessment is much more complex and important in management. In science the only risk is being thought a fool by the referee.

Results

Levels of evidence

Anecdote and casual observation

Though many scientists would rather die than admit that anecdotal evidence was "scientific"; actually many scientific hypotheses probably start life from a chance observation. Such observations establish the potential existence of the process. This serves to stimulate further work to confirm the process and investigate the effect. Of course, not all anecdotes are equally credible, a story told about an animal by a man in a pub may not inspire much confidence - unless the story is told at a conference. Stories told in the conference bar are often the start of new hypotheses and research programs.

Logic, plausible arguments and mathematical modelling

Starting from a set of assumptions (and sometimes a smidgin of data) the thinker predicts that a certain consequence is likely. This is often how people conclude: "the process exists therefore it must be important". Unfortunately the success of pure thought in science is patchy, to say the least - particularly in ecology. Data has often shown that nature has a depressing habit of being much more complicated than we expect.

There is no doubt that logic is an indispensable tool in science. But on its own, while it is a strong tool for suggesting hypotheses, it is a weak one for confirming them. Arguments that A causes B which causes C, therefore since A is present so must C, are only as good as the supporting data for the inevitability of each of the links and the absence of any alternative plausible models for the causation of C. Computer or mathematical models are merely formalised logic where the assumptions and logical sequences are explicit so the argument if-this-then-that, can be shown to be valid. It is therefore distressing to find that scientists not infrequently seem to accept logical argument and/or mathematical (or computer) models as sufficient justification for accepting a hypothesis or prediction. MacArthur and Wilson's (1967) Equilibrium Theory of Island Biogeography is an example. An elegant mathematical model suggested a plausible explanation for observed patterns (e.g., that larger islands hold more species). It was accepted as a paradigm by an influential and prolific group of ecologists during the 1970s and was assumed true by much of the ecological research of the time. It was only in the 1980s that it was finally pointed out that the model had never been tested, and that much available data actually contradicted it (Gilbert, 1980; Williamson, 1981).

Computer or mathematical models that claim to show what happens if certain assumptions are made (and all models make assumptions) can be useful in suggesting what *could* be happening in the real world. But in the absence of data they are of no use in determining what *is* happening. In particular, simple models will seldom be of much use in field situations since they generally do not allow the same richness of behaviour that the real system can show, they therefore usually will fail to suggest the full range of what could happen. As a result they are of limited use even for suggesting hypotheses. Like logical thought (which is all they are, expressed in a formal fashion) their conclusions will only be adequate if they do not leave out relevant processes or starting conditions.

In the end these models are of most use as descriptions of the relationship between the processes and the effects we observe. While those descriptions are unsupported by evidence the model remains a hypothesis, after the data have been accumulated for the various links of the model, its predictions have been tested and there are no plausible alternative models, then our model becomes a useful description of nature.

Computer and mathematical models are best used as formalised hypotheses. In this form they can be tested for internal consistency and consistency with existing data. At their best they can make predictions or have structural features that can be tested or investigated in the field allowing them to be falsified. However, as evidence they are only as strong as the data that support them, and the number of alternative models the same data would also support.

Observational data

Observational studies (as opposed to manipulative experiments) are traditionally the major source of information about natural systems. In recent years they have become a lot less respectable being derided

as 'mere natural history' and 'not real science'. This is grossly unfair. Observational data can still be a useful (sometimes the only) source of information on important topics, so long as its limitations as evidence are appreciated. It only becomes bad (or at least inefficient) science if there are practical ways of getting better information about the same subject, or if the evidence is given more weight than it can sensibly support. At its best observational data can establish the existence of a process, but usually provides only weak evidence of the magnitude and generality of the effect. Arguing from pattern to process is a dangerous thing to do, since any one pattern is rarely explicable by only one process. We can separate two types of observational evidence - one which works and one which doesn't.

Inference when there is no 'referent'

In order to demonstrate an effect we must be able to compare situations where the effect is present/strong with those where it is weak/absent. The effect is measured (and demonstrated to be non-trivial) by the difference between the two. This is the basis of correlation.

A great deal of bad blood was generated in the early eighties by an argument over inferring competition from the distribution of bird species on islands. Surveys had shown that some bird species were never found together on the same island (Diamond, 1975). One group of workers inferred that this was due to the effects of competition (e.g., Diamond and Gilpin, 1982), and another group suggested that alternative hypotheses were equally plausible (e.g., Connor and Simberloff, 1979, 1984). There were some fairly silly things said by both sides, in particular about the role of random combinations of species as null hypotheses; but ultimately it became clear that data of this kind can provide only minimal evidence in support of competition. They could not compare islands where competition had occurred with islands where it had not. So they could not show the effects of competition. In this type of situation there are simply too many alternative explanations, and the evidence will seldom be good enough to distinguish between them (Harvey and May, 1985).

A similar logic was introduced to me as an undergraduate to show that heavy parasite loads increased the death rate of the host. The number of parasites per host is often well described by a negative binomial distribution (Crofton, 1971). So, if it was observed that the negative binomial appeared to be truncated (no very heavily infested individuals in the sample), we could infer that the heavily infected individuals had died. Unfortunately, as later authors pointed out (e.g., Anderson and Gordon, 1982), there was no reason to believe that those hosts had even existed; their existence had to be inferred indirectly. Other processes including inter-parasite competition for space, host immune responses, and age specific infection rates could all lead to such truncated negative binomials. Since the heavily infected individuals were never observed, their death rate could not be compared with the rest of the population.

Inference from correlations

Let us extend the previous example and suppose that we were able to directly compare the parasite load of dying individuals with the rest of the population. In this case a difference might be due to the effects of the parasites. We have a stronger case than before. However this example again shows up the problems with this type of information. There is at least one alternative explanation: the ones that die are older, older animals have had more opportunity to accumulate parasites. It is the old problem that correlations are poor evidence for causality. Such correlations can establish the possibility of an effect, they may also allow its estimation; though we can seldom distinguish these from other causal explanations. It is worth noting that if appropriate extra information (e.g., age) has been gathered, then appropriate analytic techniques (partial correlations, multiple regression (Sokal and Rohlf, 1981), or generalised linear models (Crawley, 1993)) can be used to separate out the effects of age from those of parasite load to help distinguish between the alternative explanations.

There is a classic data set that relates the size of the human population in the city of Oldenberg between 1930 and 1936 with the number of storks nesting there in the same years (correlation coefficient = 0.92, P=0.003). Before we race off to rewrite the gynecological text books we perhaps ought to consider alternative hypotheses. Since the storks nested on the large chimney pots of rich people's houses we might wonder if the number of such chimney pots remained constant over that time. In fact the city grew quite considerably over that period so both the number of babies and the number of storks increased. I would have liked to see the figures for the nineteen fifties. The number of babies would have increased as the city continued to grow, but the number of storks might have declined due to DDT. A significant negative correlation would be almost inevitable. Would this allow us to suggest storks as contraceptives?

Examples like this occur commonly in ecology. For example, in sites where a potentially competing species B is absent, species A has a higher population size/birth rate/survival. In the past

evidence like this was accepted as demonstrating the action of competition; regrettably, in virtually every case there was insufficient information to exclude other equally plausible hypotheses like differences in habitats between sites, or shared parasite or predator species.

Separating the effect of interest from plausible alternatives will generally be difficult in natural situations - hence the interest in designed studies (see below). But so long as the appropriate information was collected and care and sound biology are applied to the analysis to remove the effect of plausible confounding factors like age or habitat, and we have large sample sizes, then some confidence can be placed in the conclusion.

Designed studies

In recent years many ecologists have realised that extracting data from non-manipulative studies is extremely difficult. They realised that to get the strongest possible evidence from a study they could use many of the strategies associated with the design of manipulative experiments. These "pseudoexperiments" as they are sometimes unkindly known have the same objectives as traditional manipulative experiments, but have the major restriction that there is no control of where and how the process operates within the study. However, since they use the same methods (to a degree) I will consider some of these methods first, and consider the differences between designed observational studies and manipulative experiments at the end.The primary aim of any such study is to measure the effect that the process has caused.

The two major aims of any design are: to separate out the effect of interest from all other effects (i.e., avoid confounding) and to allow generalisation of the effect to some defined statistical population in space and time.

Separation of effects

The chief aim of a designed study (experiment or pseudoexperiment) is, where possible, to separate out the effect of the process under study from the environmental effects that might be confused with it. For example, a study was published some years ago by two workers (who now know better) where two subtidal rocks were selected, and sea urchins were cleared from one of them, the other being left as a control. The barnacle density was recorded from both some time later and the difference between the two densities attributed to the sea urchin removal. The problem with this design is that the difference between the two densities could equally plausibly be due to other differences between the rocks as barnacle habitat. The authors had confounded (failed to separate) the effect of sea urchin clearance from the natural variability of rocks in the density of barnacles they can support. As we shall see below this is a clear case of pseudoreplication (Hurlbert, 1984).

The separation of effects is done in two main ways. First, by randomly allocating treatments over all possible combinations of other effects we make sure that on average our treatment effects are uncorrelated with these other effects and are therefore separate. The trouble is that this is on average, over many repetitions of the experiment, which is not much use for any particular experiment. There will, therefore, be correlations between the treatment effect and environmental effects. If the sample size is small then the correlations could be quite large; so within a particular experiment there could be confounding. Of course if it is obvious then we throw away that randomisation to try another. Still, some correlation will exist. This is allowed for in the significance test which is why randomised designs are of such low power -by failing to avoid this confounding they are very inefficient.

The other main way to separate effects uses orthogonal design, meaning that within the experiment the effects of the designed factors are uncorrelated. Within a particular experiment the treatment effect is separable and unconfounded. This design is clearly superior to randomisation which relies on a "over all possible experiments" to work. Because there is no possible correlation between the treatment effect and the other factors incorporated into the design, the treatment effect is much more precisely estimated. This is why blocked and repeated measures designs are so useful. (Of course we must never forget that other, possibly unrecognised, factors are operating so we must randomise over them).

Example

Let us consider the difference between a nested (random) and a randomised complete block (orthogonal) designs. The scenario is that we are to set up 3 treatments on 6 rocks with 3 cages on each rock. Unknown to us, rocks are protected (P), medium (M) or exposed (E) to wave action.

First let us consider the random design. We randomly choose 2 rocks to receive treatment 1, 2 rocks for treatment 2, and 2 rocks for treatment 3. In Table 1 are the first 7 such designs I got from my random number tables. Clearly all the designs have the treatment to a greater or lesser extent confounded with rock type. On average over repeated experiments they would not be. With larger numbers of rocks they would be less likely to be so.

Table 1: *A random nested design applied to seven random samples of rocks.*

Design	I	II	III	IV	V	VI	VII
Cages							
1,1,1	E	M	M	E	E	E	M
1,1,1	P	M	E	E	E	E	M
2,2,2	E	E	P	M	P	P	E
2,2,2	E	M	M	E	P	M	P
3,3,3	M	M	P	E	M	P	E
3,3,3	M	M	M	M	M	M	E

Remember we do not actually know which rocks are exposed. We would have to accept any one of these designs blindly.

There are clearly two problems. The first is that the sample of rocks chosen was sometimes not representative of the population as a whole (e.g., Design II, perhaps also Design IV). The second problem is that there was confounding of the treatment effect with rock types, so most of the designs were unacceptable (I, II, IV, V, VI, and VII). The significance test will (on average, over many parallel universes) have the correct probability, but this does not mean that our particular experiment has a 0.05 probability of a type I error (a false claim of significance). For example design V is certain to give us a 'significant result' even if the treatments have no effects at all.

Solution

We must ensure every treatment is equally affected by the rock differences. They will therefore need to be uncorrelated with the rock differences. Every treatment must apply on every rock. In Table 2 a design of this kind is shown applied to a number of random samples of rocks.Now we can estimate the effect of each treatment for a random sample of rocks and get their average. Because the treatment effects are now orthogonal to the rock effects there is no confounding with the comparison between the treatments and because there are more treatment*rock combinations we have a smaller standard error for each treatment effect (we measure the effect on a larger sample from the population of interest - rocks). We still have the problem that a small sample may not be representative of the population as a whole e.g., designs II and IV, but at least we now have a better estimate of the treatment effect.

Table 2: *A randomised complete block (orthogonal) design applied to 7 random samples of rocks.*

Design	I	II	III	IV	V	VI	VII
Treats							
1,2,3	E	M	M	E	E	E	M
1,2,3	P	M	E	E	E	E	M
1,2,3	E	E	P	M	P	P	E
1,2,3	E	M	M	E	P	M	P
1,2,3	M	M	P	E	M	P	E
1,2,3	M	M	M	M	M	M	E

This design does make one assumption, that there are no slopping-over effects: the presence of a high density cage does not influence the effect in a low density cage on the same rock.

This kind of design can be difficult to achieve in observational studies, but a similar approach can sometimes be used if the appropriate information is available. For example, suppose that we wish to investigate the effect of predation by sea urchins on barnacles but because we are working in a nature reserve we are not allowed to manipulate their densities. Now some rocks have urchins and others do not. We first need to identify and measure any environmental factor that could be correlated with the density of urchins; if they also influence the density of barnacles then they could be confounded with the effects of urchins. Let us suggest that wave exposure is such a variable. We can reduce its influence on the urchin effect by choosing one of each type of rock with urchins on and one of each type without urchins. The main problems with this approach are:

(a) there are usually rather a large number of possible confounding variables so a design that accommodated all of them would tend to need an awful lot of rocks. Even measuring some of the other important environmental variables and correcting for them in the analysis requires more work and more rocks.

(b) If known variables don't confound the design, then unrecognised ones will. Because in an observational study we cannot randomise over the unknown variables, we do not even have this protection. There may be some unknown variable that affects both urchins and barnacles. While designing observational studies using experimental design principles clearly can improve the strength of the evidence we can get, it will still seldom give as clear results as a manipulative experiment.

Generalisation to defined population

For the results of a study to be interesting they must imply something about a wider range of situations than were studied. People will usually not be interested unless they can accept that the same processes will operate in much the same way in other places and at other times. To let them know how reliable that extrapolation is likely to be we

must define the statistical population to which the results can be applied.

It seems fairly obvious that if we want to generalise to a population we should have a representative sample from that population. For example, to formally generalise the results of a study to all kauri forests in New Zealand we should have a representative sample of kauri forests, not just one. This formal generalisation is made by using the variability of forests as our error in any analysis. We shall see below that such generalisations can be made informally, making use of data external to the current study, so that even when a study only looked at one kauri forest some generalisation may be possible.

Designed observational studies versus manipulative experiments

Observational studies

Because the study design will often be unable to separate the effects of interest from the confounding environmental factors, very often these will have to be separated in the analysis. This puts a strong reliance on techniques of statistical modelling like partial correlation, multiple regression, and generalised linear models. While these are the indispensable tools of the observational ecologist they are not simple to use properly. Also since these procedures rely on the biological appropriateness of the statistical model (with all its assumptions), they can seldom provide totally convincing evidence of an interaction's influence. However they can materially increase the strength of the evidence we can get from observational data.

Manipulative experiments

In recent years it seems to have become fashionable to assume that only a manipulative experiment can give 'scientific' results. A well designed experiment can indeed give more rigorous results than virtually any other form of investigation. But because the manipulation is by definition unnatural, the results of the experiment are to that degree less relevant to what is really happening. For example, in the experiment the treatment will be applied to randomly chosen individuals. In nature most effects are not randomly distributed over sampling units. As a result of this artificial mimicking of nature we may achieve an unambiguous, precisely measured effect of some manipulation, but because the change that was made was deliberately separated from the correlated changes that might accompany it in nature, the effect may not be the naturally occurring effect. The experiment merely establishes what could occur if the change mimicked by the experimental treatment happened naturally. It will seldom be able to show what would happen. Much will depend on the existence of appropriate controls. Experiments tend to gain in rigour and precision as they lose in relevance (at the extreme they are done in the lab). A badly designed experiment will usually produce weaker evidence than a good observational study.

Pseudoreplication

Having looked at the role design principles can have in planning ecological studies it seems appropriate to look at one of the most influential concepts to enter ecology in recent years - pseudoreplication (Hurlbert, 1984). It has probably been used to reject more field ecology manuscripts than virtually any other reason, but there is ample evidence that many workers do not understand the limitations and implications of the concept. It is defined as " the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated or replicates are not statistically independent" (Hurlbert, 1984). Part of the problem is that the definition conflates two quite separate issues: the purpose of replication and the independence of the sampling units.

Purpose of replication

Replication gives the ability to separate the treatment effects from other effects. Clearly this requires us to replicate the treatment on more than one sampling unit so that the treatment effect can be separated from the sampling unit variation. This then is the true root of pseudoreplication as a problem of design, the necessity to replicate appropriately in order to separate effects from one another.

This aspect of pseudoreplication also has implications for the analysis. As was mentioned above, in order to generalise formally to a population, we must have a representative sample (i.e., more than one unit) from that population. In particular we must have an error term derived from that random sample. This is the core of the confusion over the choice of error term in ANOVA. If we wish to generalise to a population from which a random sample was taken, like the rocks in the nested design of Table 1 then we must use the variability associated with the rocks to make inferences. If we wish to consider the rocks we used as the complete population (not a sample), perhaps because they were all the rocks that were in the bay, then there is no rock error and we would use the within-rock between-cage error. Of course, since such results lack generality, an audience may feel that the results are uninteresting; but we are only guilty of pseudoreplication if we claim a generality that the analysis was not designed to provide.

Independence of the sampling units
The assertion that sampling units must be independent is completely wrong. This is a fallacy largely perpetuated by word of mouth and introductory statistics courses (I have been guilty of it myself). In fact the requirement of the usual statistical methods is that the **errors** after fitting the statistical model implicit in the method of analysis (be it regression, t-test, ANOVA, or contingency table) are independent. Non-independent replicates can be validly analysed, provided the non-independence can be incorporated explicitly into the model. Nested analysis of variance has non-independent replicates, the lowest level replicates (the cages in Table 1) are correlated when they come from the same higher level unit (rocks in Table 1). This correlation means that a simple one-way ANOVA using just treatments and cages would be inappropriate, the errors would be correlated. The structure imposed by the rocks is simply incorporated into the model as the nested term, the errors are now independent even though the cages are not. Many common forms of non-independence can be incorporated simply into mixed model analyses like repeated measures and random blocking of factors. Even non-standard ones can often be accommodated by modern techniques (e.g., Linear Mixed Models (Searle, Casella and McCulloch, 1992), Generalised Estimating Equations (Waclawiw and Liang, 1993), Generalised Linear Mixed Models (Breslow and Clayton, 1993). There is little doubt that independence of sampling units can lead to a simpler analysis but often at the cost of power and cost effectiveness. Non-independence can often be exploited to improve efficiency both in sampling (e.g., McArdle and Blackwell, 1988) or significance testing (e.g., Legendre and McArdle, *in press*).

Interpretation of evidence

The strength of evidence associated with a study is not necessarily a good measure of the effect it will have on its target audience. Ecological studies do not stand on their own, they are interpreted in the light of information held in the heads of the audience. We can identify four main features of the process.

Acceptance of the results as interesting
Conclusions based on weak evidence can be interesting if the topic is of consuming interest to the audience and/or because it has proved impossible to get better data. Much of the modelling work in population and community ecology got published on this basis. These studies contain little more than formal speculation but in a field where there is little better available they were of interest and use.

Acceptance of the existence of an effect
If the evidence from the current study contradicts strong prior beliefs, it is less likely to be accepted, even published. Naturally enough the prior evidence can be classified in the same way that I have used in this paper. We can distinguish between prior beliefs based on data, theory, intuition, or prejudice. If the current study contradicts strong data-based evidence from previous studies it is unlikely to be accepted unless the contradiction can be reconciled or the evidence in the current study is perceived as more compelling. If the current study contradicts the current theories or models it can have difficulty in being accepted. But clearly its acceptance will be more likely if the results generate new plausible models of their own.

Counter intuitive results usually require stronger evidence for acceptance. Beliefs based on prejudice are clearly the most difficult to overcome. Generally the strength of the study's evidence has little to do with this.

Acceptance of the generality of the effect
While we addressed earlier the problem of formally generalising the results of a study, most scientists use a sort of informal generalising based on their knowledge (often intuitive) of the system being studied. For example many physiologists will take measurements from a single individual and then generalise to all the members of the same species. If the character being measured is known to be relatively invariant in related species, like the oxygen binding capacity of the blood, such a generalisation is quite defensible. If the character is known to be variable then such a generalisation would not be acceptable. In fact this informal generalisation is widely practised and is the basis of most scientific inference. Those workers who demand formal inference or nothing are ignoring the successful application of informal generalisation by generations of scientists. In ecology the problem with this approach is that many of the effects being studied are clearly variable, often extremely, in space and time. Thus if we wish readers to generalise our results, it is probably necessary to provide arguments as to why a study can be taken as representative of a wider class of situations.

Acceptance of the importance of the effect
Once the existence and magnitude of an ecological effect has been accepted it is quite another thing to show that it has consequences at the level we are interested in. For example, we may have shown that there is a measurable death rate due to predation; this does not mean, however, that this will have any measurable effect on the real population growth rate.

I can still remember an argument where I refused to believe that an over 90% larval death rate due to a parasitoid was not limiting a wild population of *Drosophila*. In fact, as I finally appreciated, the number of individuals in the next population was entirely determined by the number of windfall apples in the orchard. Such was the fecundity of a single female that only a relatively small percentage of the population were able to breed anyway. Reducing the death rate would have no effect on the number of larvae in the next generation. As a population factor this enormous death rate was virtually irrelevant.

A further complication is that ecological processes operate at a number of temporal and spatial scales. It is often difficult to show that a process, though having a major effect at one scale may be having little or no effect at another, more interesting, scale.

Conclusions

The correct assessment of the quality of evidence is at the root of good science. This is not to say good science requires only strong evidence, that would be nice but we can progress, albeit slowly and carefully, with less adequate information. For example, managers may be required to act even in the absence of good information because risks and costs of waiting more than outweigh the costs of making a wrong choice based on poor evidence. Even in science, low quality evidence can be useful in the sense that it is publishable - the apparently universal criterion. In the absence of any information about a subject, if the subject is interesting, scientists will accept gratefully even poor quality information - it is better than nothing. So long as their deficiencies are born in mind, poor quality results may serve to generate new hypotheses or modify our perception of existing ones. Physiologists and molecular scientists regularly publish results on one animal, animal behaviourists publish results from incredibly artificial laboratory experiments - their relationship to reality (what animals actually do rather than what they can do) is often small but other scientists presumably find these results useful and interesting.

So, we can use weak evidence about agents of decline in the New Zealand biota if we are careful to never claim more than it can bear. One thing we should avoid is the cry I hear all too often when I point out that the evidence will not support an assertion: "But it's so hard to get good data!". Yes, such scientists have my sympathies. Nature is awkward that way, and that means their information may well be that much more interesting to their co-workers because something is better than nothing. But it does not change the quality of the evidence. If it's weak it stays weak, and conclusions based on it must be appropriately worded.

The final assessment of the strength of a study about competition or predation or disease will depend on other evidence already available, this allows the gradual accumulation of evidence in the same way that lawyers will "build a case" out of largely circumstantial evidence. This approach can also work in science but we have to use great care that we emphasise not the weight of the accumulated evidence but its ability to distinguish between competing hypotheses. An inconclusive study repeated a number of times may produce a lot of data, but will be no more useful than the same study done once if it cannot help us choose between alternative models. In the end, the main function of evidence, however accumulated, is to help us identify the "true" model for a situation. If it is not good enough to do that, then wisdom forces us to say the most difficult words for any scientist: "I don't know".

References

Anderson, R.M.; Gordon, G.M. 1982. Processes influencing the distribution of parasite numbers within host populations with special emphasis on parasite-induced host mortality. *Parasitology 85:* 373-398.

Breslow, N.E.; Clayton, D.G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88:* 9-25.

Connor, E.F.; Simberloff, D. 1979. The assembly of species communities: Chance or competition. *Ecology 60:* 1132-1140.

Connor, E.F.; Simberloff, D. 1984. Neutral models of species' co-occurrence patterns. *In:* Strong, D.R.; Simberloff, D.; Abele, L.G.; Thistle, A.B. (Editors), *Ecological Communities: Conceptual Issues and the Evidence*, Princeton University Press, Princeton, New Jersey, U.S.A. 613pp.

Crawley, M.J. 1993. *GLIM for Ecologists*. Blackwell Scientific Publications, Oxford, U.K. 379 pp.

Crofton, H.D. 1971. A quantitative approach to parasitism, *Parasitology 62:* 179-193.

Diamond, J. M. 1975. Assembly of species communities. *In:* Cody, M.L.; Diamond, J.M. (Editors), *Ecology and Evolution of Communities,* pp. 342-444. Harvard University Press, Cambridge, Mass., USA.

Diamond, J.M.; Gilpin, M.E. 1982. Examination of the 'null' model of Connor and Simberloff for species co-occurrences on islands. *Oecologia 52:* 64-74.

Gilbert, F.S. 1980. The equilibrium theory of island biogeography: Fact or Fiction? *Journal of Biogeography 7:* 209-235.

Harvey, P.H.; May, R.M. 1985. Competition in imaginary worlds. *Nature 314:* 228-229.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs 54:* 187-211.

Legendre, P.; McArdle, B.H. The comparison of surfaces. *Oceanologica Acta,* in press.

MacArthur, R.H.; Wilson, E.O. 1967. *The Theory of Island Biogeography*. Princeton University Press, Princeton, New Jersey, U.S.A. 422 pp.

McArdle, B.H.; Blackwell, R.G. 1989. Measurement of density variability in the bivalve *Chione stutchburyi* using spatial autocorrelation. *Marine Ecology, Progress Series 52:* 245-252.

Searle, S.R.; Casella, G.; McCulloch, C.E. 1992. *Variance Components*. John Wiley and Sons, New York, U.S.A. 501 pp.

Sokal, R.R.; Rohlf, F.J. 1981. *Biometry*. 2nd ed., W.H. Freeman & Co, San Francisco, U.S.A. 859 pp.

Waclawiw, M.A.; Liang, K-Y. 1993. Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association 88:* 171-178.

Williamson, M.H. 1981. *Island Populations*. Oxford University Press, Oxford, U.K. 286 pp.