Introduction to statistical analysis of invertebrate monitoring data

Version 1.0

This specification was prepared by Alison Evans in 2016.

Contents

Disclaimer
This document contains supporting material for the Inventory and Monitoring Toolbox, which contains DOC's biodiversity inventory and monitoring standards. It is being made available to external groups and organisations to demonstrate current departmental best practice. DOC has used its best endeavours to ensure the accuracy of the information at the date of publication. As these standards have been prepared for the use of DOC staff, other users may require authorisation or caveats may apply. Any use by members of the public is at their own risk and DOC disclaims any liability that may arise from its use. For further information, please email biodiversitymonitoring@doc.govt.nz

Department of Conservation
Te Papa Atawhai

Summarising the data

In this section, examples of how to summarise and display data are provided, and where possible, worked examples are provided of the appropriate statistical analysis. This section is an outline of the likely types of analysis used for invertebrate conservation studies rather than an exhaustive list. The analysis of each study will depend on the study design, type of data and the hypothesis being tested rather than on the trapping methods used. It is strongly recommended that statistical advice is sought early to check that the study design is appropriate and that you will have enough samples to conduct meaningful statistical analysis. Once the data have been collected, they will need to be checked carefully for specific one-off errors (e.g. writing the same plot number down twice when collecting the data) and for systematic errors (e.g. taking too few samples each time) which limit the statistical power of the data. Once the data have been summarised using tables and graphs, it is sensible to again seek formal statistical advice to help with the analysis. The most common way to summarise the data is to use a spreadsheet such as those available in Excel or R. It is important not to combine any of the data points (e.g. lumping all of the data from a grid of pitfall traps) until you have had advice from a statistician, as it is sometimes difficult to split data retrospectively. Your statistician may ask for the data to be set out in a specific way on the spreadsheet and it saves time (and money) if you can prepare the data accurately, ready for analysis. An example of a properly formatted data sheet for a replicated study is provided in Appendix A. Remember not to indicate missing data or blank cells with zero (as zero is a valid data point). A key is provided in Appendix B to help determine which statistical test to use (Crawley 2002).

Although the examples provided are based on real projects undertaken by DOC or other researchers, some of the scenarios and the data have been manipulated to demonstrate the analysis. It is therefore NOT appropriate to cite these studies or the data in other documents unless they are accompanied by references.

Calculating sample size (*n*)

The sample size will be used extensively in your calculations. Sample size (*n*) is the number of independent sampling units (e.g. number of traps or the number of treatments if it is a replicated study). For example, if you wanted to see whether the abundance of beetles differs between two sites with 15 pitfall traps in each area, the sample size would be *n* = 15 in each site.

Arithmetic mean

The mean value is sometimes referred to as the average. Mean values are the basic building blocks of your data summary and statistical analysis. The mean can be calculated by summing the number of individuals and dividing by the sample size (or number of observations).

$$\chi = \sum (x1 + x2 + x3 \ldots.)/n$$

For example, the mean number of snails observed in quadrats on one occasion can be calculated as follows.

$$\chi = \sum (4, 5, 2, 0, 5, 7, 21, 11, 4, 6, 6, 3, 7, 4, 5)/15$$

$$\chi = 6$$

Median

The median is a measure of central tendency used to describe the middle value when the data is put into rank order. For example, the median value for the snail data presented is calculated as follows: 0, 2, 3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 11, 21. In this case there are two values in the middle (5, 5). In this case you add the two values together (5 + 5) and divide by 2 (10 ÷ 2 = 5). NB: If your data is normally distributed the mean value will be similar to the median value.

Range

This is simply the values that the data lie between. For example, the range of the snail data above (4, 5, 2, 0, 5, 7, 21, 11, 4, 6, 6, 3, 7, 4, 5) is 0–21. The range is often given in association with mean values and gives an indication of how variable the data is. See table 5 in Eason et al. (1993) if you would like to see an example of how the range is presented as part of a statistical analysis. The Excel Analysis Toolpak gives the range as a single figure (the difference between the highest and lowest number). In our example this would be 21.

Distribution of the data

Before deciding on what type of statistical tests are available to you, it is important to know whether your data is normally distributed and follows a typical bell-shaped curve (Fig. 1).

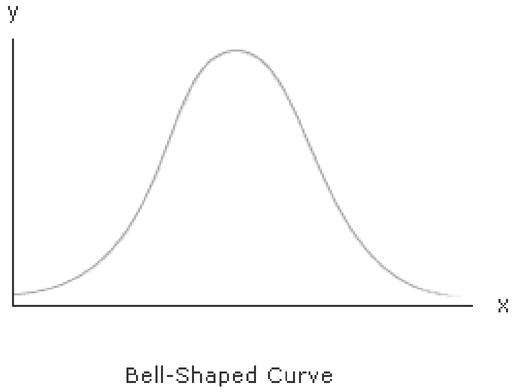

Figure 1. The distribution of data that fits a normal or bell-shaped curve where 95% of the observations fall within 1.96 standard deviations of the mean.

Some statistical tests are better suited to normally distributed data or data that can be normalised by transformation (parametric tests) and some are better suited to skewed data (nonparametric tests). Your statistician will be able use a 'goodness of fit' test to determine whether your data are normally distributed and which test is most appropriate. The first step in determining whether your data are normally distributed is to graph them as a histogram so you can immediately see if they have an approximate bell-shaped curve or not. Count data (whole numbers, such as number of trees in an area) will follow a Poisson distribution (very skewed with lots of zeros) if counts are mostly small (0, 1, 2). Here the mean is approximately equal to the standard deviation. The standard deviation describes how far the data spread from the measure of central tendency. It is commonly used to describe data as it is presented in the same units as the original measurements. When the data are normally distributed, 95% of the data (or observations) will fall within 1.96 standard deviations of the mean leaving 2.5% on the left side of the bell shaped curve and 2.5% on the right side of the bell shaped curve.

Useful tips on understanding standard deviation are available at http://statistics-help-for-students.com/What_are_measures_of_central_tendency_and_dispersion.htm#.VrPDQLJ97b0

With larger counts the distribution becomes binomial (less skewed) and if counts are sufficiently large (lots of counts in the range of 10 or more) the binomial distribution starts approaching the normal distribution. There are specific statistical methods for treating data from each of these distributions. In general, the larger your sample size the more confident you can be that your data will be normally distributed and that you can use parametric tests. As a 'rule of thumb' if your sample size is less than 30 then the distribution of your data becomes more important.

Displaying summary data (tables, histograms, box and whisker plots, scatter plots)

Explore the data you have collected using graphs to first identify patterns, trends and potential relationships before you start using a statistical test. It is also important to summarise the mean values and provide an indication of the variability within the data. The most common way to summarise data is in a table or in a graph using a program such as Excel or R. In Excel you can get an overall summary of your data by going to: 'Tools > Data Analysis > Descriptive Statistics'. This also provides a summary for each column (or row) of your data including mean, median, mode, standard deviation, range, maximum and minimum, and 95% confidence intervals. Histograms showing mean values for each invertebrate group are a good place to start. You can add error bars (or 95% confidence intervals) to each of the mean values. Large error bars indicate that there is high variability in your data and that your sample size (number of traps) might be too small to provide meaningful data. The degree of overlap between the error bars can also give an indication of whether there is a significant difference between your mean values. If the standard error bars overlap, the means are not likely to be significantly different at $p < 0.05$ level.

In the example provided (Fig. 2), there is overlap between the error bars at distances of 10–70 m suggesting that there is no significant difference in katipō numbers recorded at these distances. Similarly there is no overlap in the error bars between katipō numbers recorded between 80–100 m from the shoreline. A summary statement about the data would include a comment that there appears to be significantly fewer katipō recorded at distances > 70 m from the shore. A statistical test can confirm this for you.

NB: If the data are normally distributed (see Fig. 1) and the standard error bars overlap then this means that there is no significant difference between the means, but if the data are not normally distributed then *either* a more sensitive test is required *or* a transformation should be applied to the data to normalise them before they are graphed.

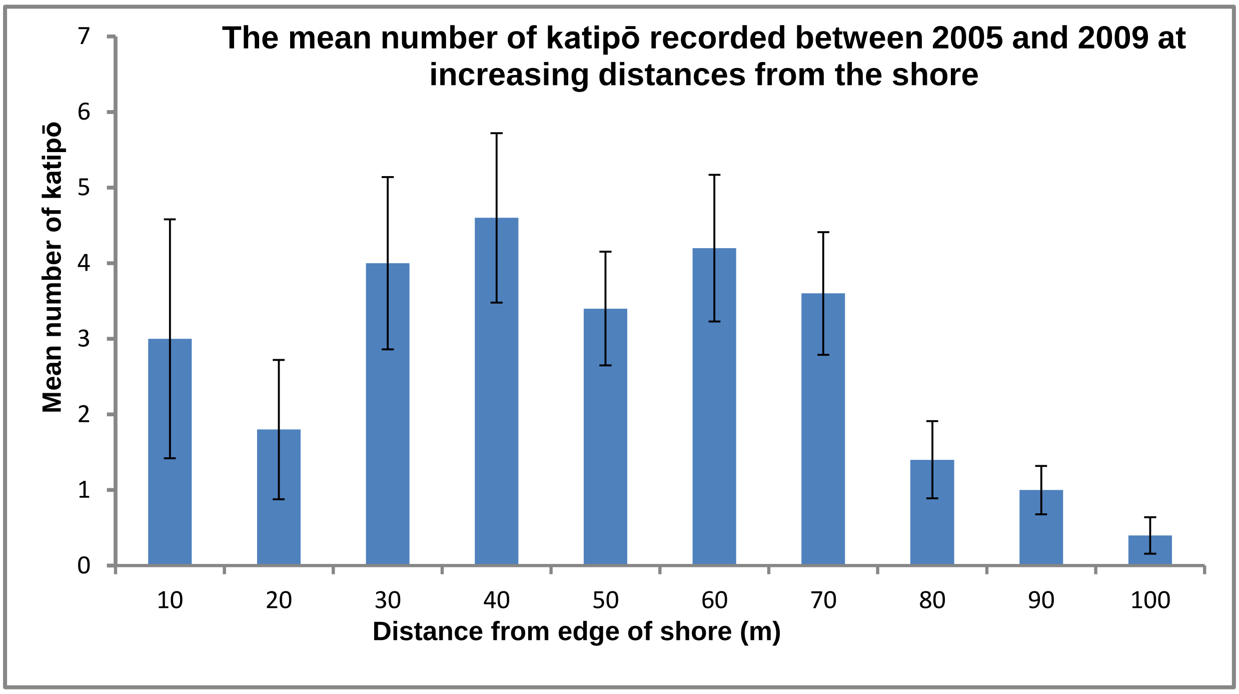

Figure 2. The mean number of katipō spiders recorded at increasing distances from the shoreline over a 5-year period at Kaitorete Spit in Canterbury, New Zealand (I = standard error).

Line graphs are a good way to present data showing changes over time (temporal patterns). You can add error bars and 95% confidence intervals to this type of graph using Excel. A good example of this type of data is given in Powlesland et al. (2005).

In the following example, Flax snails *(Placostylus hongii)* were monitored over a 13-year period between 1996 and 2008 at Peach Cove in Northland. Ten randomly chosen transects, each 25 m long, were placed through the snail habitat. Circular plots (1 m diameter) were placed at 5 m intervals along each transect to assess snail abundance. The results are presented as a line graph with standard errors associated with each data point shown (Fig. 3).

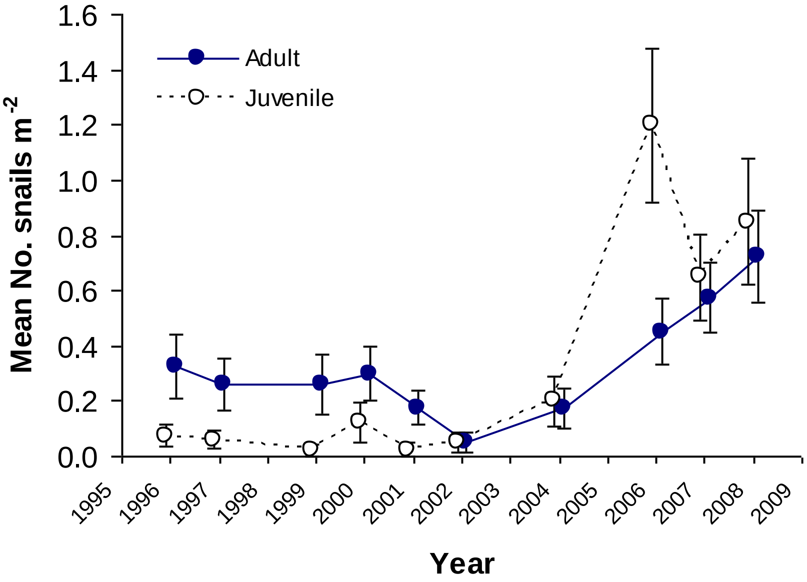

Figure 3. Change in density of live *Placostylus hongii* snails at Peach Cove. Standard error bars are shown. Adult and juvenile results are jittered (offset) for clarity.

Box and whisker plots are another way of graphically representing your data. They are more suited to data with few variables but give an indication of the interquartile range of the data (the box showing the upper quartile above which 25% of the data lie and the lower quartile above which 75% of the data lie) and the standard error (the whiskers). Good programs to produce this type of chart are R or SPSS.

Many studies focus on the beetle fauna since they represent a range of trophic levels, they are better known taxonomically than other groups and thought to reflect the diversity of other invertebrate groups (Hutcheson et al. 1999). Data collected from inventory studies can be presented so that the beetle fauna is divided into three main trophic levels: predators, herbivores and detritivores (these may be divided further depending on the scope of the study). This gives coarse information about the beetle community composition and the complexity of the habitat that they were collected from. In the example provided (Fig. 4) the beetle community has been divided into four trophic levels and compared between pasture, young kānuka, old kānuka and mature forest (Dugdale & Hutcheson 1997).

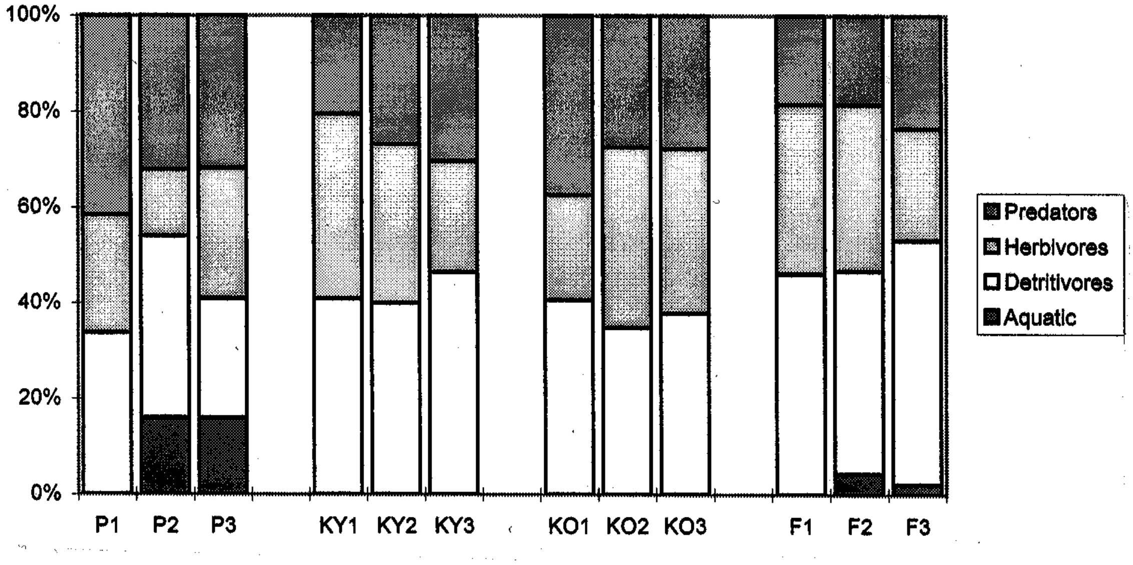

Figure 4. The percentage functional composition of the coleopteran communities occupying pasture (P1, P2, P3), young kānuka (KY1, KY2, KY3), old kānuka (KO1, KO2, KO3) and three mature forest sites (F1, F2, F3) in the Gisborne region (Dugdale & Hutcheson 1997).

Scatter plots offer a diagrammatic way of displaying relationships between data, especially if you suspect there is a cause and effect relationship. For example, there may be reason to suspect a relationship between distance from a forest edge and temperature in the following scatter plot (Fig. 5).

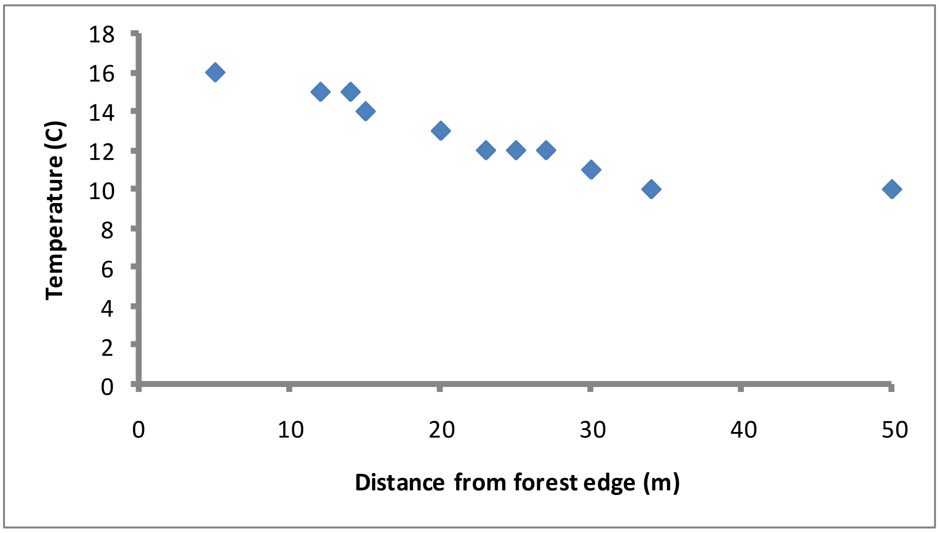

Figure 5. The relationship between distance and temperature recorded at increasing distances from the edge of a beech forest in January 2007 in the Lewis Pass.

It is also possible to fit a trend line to a scatter plot but this is usually done when examining a relationship using regression analysis.

P-value

A *P*-value is an estimate of the probability that a result could have occurred by chance, e.g. a 1 in 20 chance of happening is *P* = 0.05 (i.e. 1/20) and a 1 in 100 chance of happening is *P* = 0.01 (1/100).

Significance, *p*-values and errors

In statistics, ‘significant’ means that a result is unlikely to have occurred by chance. The general agreement of ‘unlikely’ is that an event is unlikely if it happens less than 5% of the time—i.e. less than 1 chance in 20. Let us say your test result was *P* = 0.02. This means there is a 2% chance that the outcome of the experiment could have occurred by chance so therefore the null hypothesis (that there is no difference between the means) is rejected. We state this as ‘there was a significant difference between the means’. If the test result was, say, *P* = 0.12 then (as this is greater than *P* = 0.05) the test showed no significant difference.

Type I and Type II errors

Remember that significance is about chance. A test result of *P* = 0.05 still means that there is a 1 in 20 chance of making an error or accepting the null hypothesis when it is in fact wrong. The downside to this is that as you make it less likely to make an error in accepting the null hypothesis (by making the threshold of acceptance of the *p*-value smaller) you make it more likely that you will make an error in the reverse direction—in other words, by accepting the null hypothesis when the true situation is false. If we happen to know if a null hypothesis was true or false, then the errors we can make are shown below.

Null hypothesis	**Actually true**	**Actually false**
Accept	Correct decision	Error (Type II)
Reject	Error (Type I)	Correct decision

Variability in the data (SE and/or 95% CI)

The degree of variability in the data is considered to be a measure of how reliable or robust the data is. Studies with small sample sizes tend to show more variability and the results are therefore less reliable. The variability is most commonly reported in association with the mean. The two most common measures are the standard error (SE) and 95% confidence intervals (95% CI). Both indicate the dispersion of a distribution or the range of values that your means lie within. They can be easily calculated using an Excel spreadsheet (NB: Excel refers to the mean as ‘average’) or using R. The standard error of the mean can be calculated using the standard deviation as follows:

$SE = \frac{\sigma}{\sqrt{n}}$ where σ = standard deviation and n = sample size

The 95% confidence interval is calculated from the standard error of the mean and is considered to be a better indication of the robustness of your data. You can be 95% certain that the mean lies between the two values generated in the calculation. It is calculated as follows when the sample size is greater than about 30:

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \text{ where } \sigma = \text{standard deviation, } n = \text{sample size, } \bar{x} = \text{mean}$$

NOTE: Do not use the function '=CONFIDENCE()' in Excel—it does not produce a true confidence interval. You can obtain a true 95% confidence interval by going to 'Tools > Data Analysis > Descriptive Statistics' and ticking the box 'Confidence Interval for Mean'.

Species richness and diversity

Both species richness and diversity are often used in ecological studies, such as assessing ecosystem health.

Species richness is simply the actual number of different species (or varieties) in your collection. Diversity is a measure of the number of different species together with the 'evenness' with which they are distributed (their relative abundance). Invertebrate communities are typically very species rich or diverse and most sampling methods will provide a vast number of different species or types. There are several key things that you need to keep in mind when sampling invertebrates and assessing invertebrate data:

- You may not be able to identify all of the invertebrates to the same taxonomic level. Your analysis will assume that all individuals assigned to a specific taxonomic group are equal and interpretation of the data will need to accommodate this.
- You will often collect many juvenile specimens and because they are sometimes difficult to identify, they are often omitted from studies assessing species richness and diversity.
- Invertebrate communities are characterised by having a few common species and many different rare species. While it is important to include the rare species in your assessment of species richness and diversity, you may need to omit the rarest species in some data analyses (because the numbers will be too few to be meaningful).
- Species richness increases with sample size. It is possible to determine whether your sample size is adequate using a technique called rarefaction.

Calculating and summarising species richness data

In its most simple form species richness is simply the total number of different species that you collected. Often it is not possible to identify everything to species level, in which case it is recommended that you summarise your data at the family level or use morphospecies (species that look the same, e.g. species A). These are sometimes known as recognisable taxonomic units (RTUs). It may be possible to summarise specific invertebrate groups separately (e.g. you may have all of the beetles identified to species level). Using tables to summarise the data or using a graph are both good ways of presenting invertebrate species richness. Be mindful that your sampling method has only collected a small proportion of the species that will be occupying the habitat.

Calculating and summarising diversity data

There are many ways of measuring ecological diversity but it is often calculated by using a diversity index. This enables the researcher to describe the invertebrate community mathematically (usually with a single figure) and compare diversity between different habitats or measure changes in diversity over time. One of the more common measures of diversity is the Shannon index of diversity. The assumptions that are made by this index allow it to be used in ANOVA (analysis of variance) tests. In the example provided below (Table 1), moths were collected using a sweep net from tussock grassland near Burkes Pass, South Island. The table shows how the Shannon index can be used to calculate the diversity of moths.

Table 1. The diversity of moths (Shannon index of diversity) collected in a tussock grassland site, Burkes Pass, South Island (‘ln’ is shorthand for $\log_e$).

Moth species	**Abundance**	p_i	$p_i \ln p_i$
Orocrambus cyclopicus	96	0.357	−0.368
Orocrambus flexuosellus	72	0.268	−0.353
Kiwaia lithodes	21	0.078	−0.199
Capua semiferana	17	0.063	−0.175
Plutella psammochroa	15	0.056	−0.161
Orocrambus vittellus	14	0.052	−0.154
Eudonia cataxesta	9	0.033	−0.114
Caloptilia elaeas	9	0.033	−0.114
Eudonia feredayi	8	0.030	−0.105
Diasemia grammalis	4	0.015	−0.063
Metacrias strategica	1	0.004	−0.021
Scoparia niphospora	1	0.004	−0.021
Merophyas leueaniana	1	0.004	−0.021
Phaeosaces apocrypta	1	0.004	−0.021
	269	**1.000**	**−1.887**

Where the number of species (S) = 14

Total abundance (N) = 269

Where p_i is the proportional abundance of the ith species = (n_i/N)

Shannon diversity index is $H' = -\sum p_i \ln p_i$

$= 1.887$

A possible scenario could be that you wish to examine whether the diversity of invertebrates within forest in mainland islands differs from that in forest outside mainland islands. In this case it would be sensible to replicate the study by assessing the invertebrate fauna in a minimum of three mainland island sites and three sites adjacent to each mainland island (control). Providing care was taken to keep the sampling methods the same and the size of each sampling method constant at each site (collect the invertebrates under similar weather conditions, time of the year, similar amounts of time were spent collecting) this would allow you to compare the invertebrate fauna using a statistical test. A histogram of such data would look similar to that presented in Fig. 6.

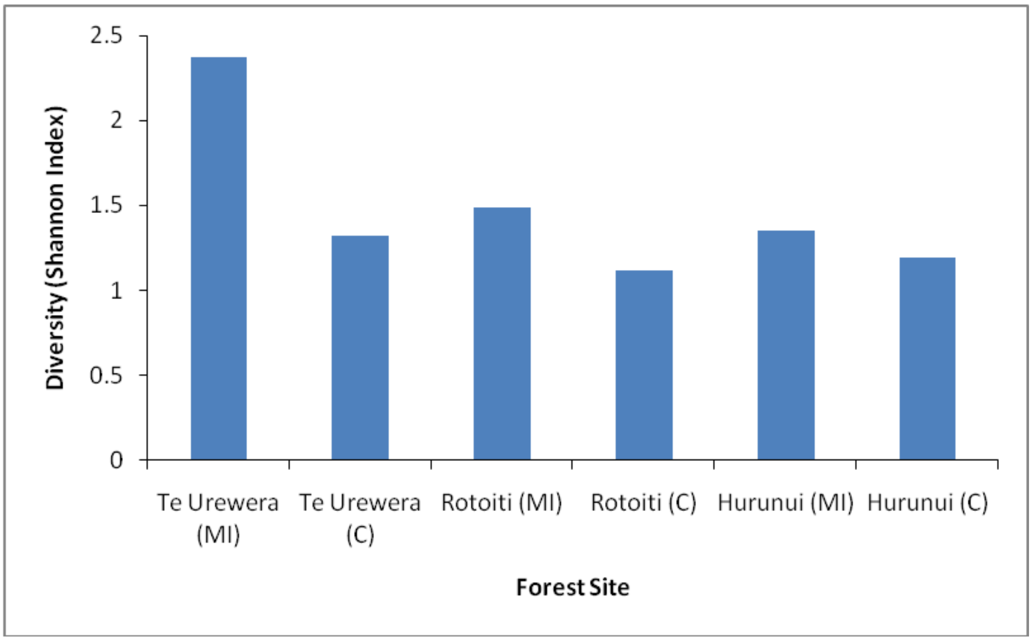

Figure 6. The diversity of invertebrates (calculated using the Shannon index) in three mainland island sites and three adjacent control sites in November, 2003.

A good guide to using diversity indices with worked examples is provided by Magurran (1988). There are many software programs available that summarise and calculate species richness and diversity. Species Diversity and Richness is a software package that is easy to follow. Otherwise programs such as R (The R Project for Statistical Computing)[1] or BioDiversity Pro[2] software developed by McAleece (1997) may be used. Advice from a statistician will be required to select the appropriate model or analysis for your data.

Rarefaction or species accumulation curves

Additional sampling of an invertebrate community will nearly always result in an increase in the number of invertebrate species that you collect. Suppose that you have 10 samples in each of your replicates or study areas. A Monte Carlo randomisation (a procedure available in many statistical packages) can be used to randomly reorder your data set, and if this is done repeatedly it will demonstrate that the first samples collected will contain a majority of the species you have collected

[1] http://www.r-project.org/

[2] http://biodiversity-pro.software.informer.com/2.0/

and that any additional samples collected will only add a few additional species (usually the rare species). The procedure can be used to determine at what point you have collected a majority of the species in a community. In the following example, a Monte Carlo randomisation test could be used to randomly reorder the invertebrate data 500 times, and then this could be used to calculate a mean value for the number of families collected for each sample size.

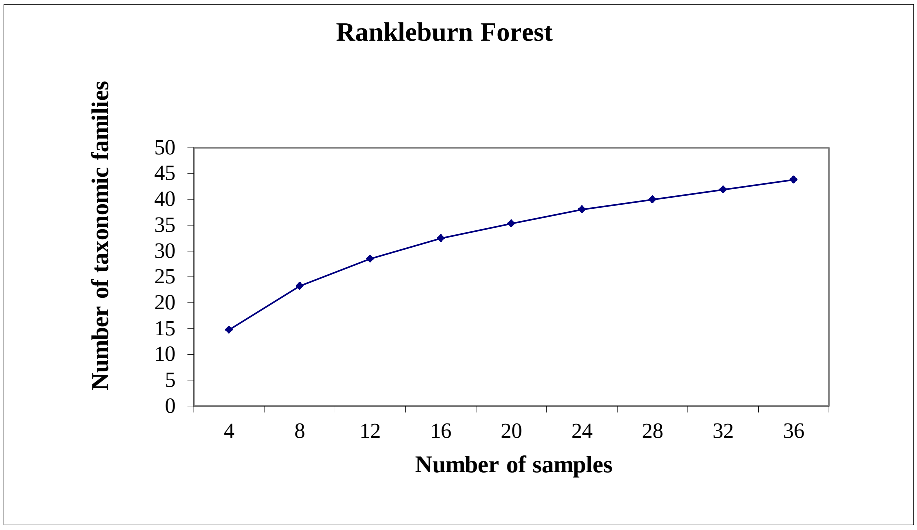

Figure 7. A rarefaction curve showing the increase in number of taxonomic families collected in Rankleburn Forest, Southland, New Zealand, as the number of samples increases.

Figure 7 indicates that as you increase the number of samples, the number of taxonomic families collected increases. The curve is slow to reach an asymptote (where the curve levels off) suggesting that more samples would be required to collect a representative indication of the number of different invertebrates occupying this forest. This test can used to predict what number of invertebrate families you would collect for any given number of samples (or traps). If you were able to identify all specimens to species level, you would expect to need even more samples before an asymptotic curve was apparent. Rarefaction may overestimate the number of species in a community because the equation makes the assumption that the invertebrate fauna is randomly dispersed when more often it is aggregated or clumped. Rarefaction is best completed following a pilot study to determine the number of samples or traps that are required to adequately sample the invertebrate fauna. The test can also be done retrospectively to determine what proportion of the invertebrate fauna you might have sampled with the number of traps you have used.

Important points to raise with a statistician

- Which parts of your data are lacking in statistical power and can this be remedied by grouping aspects of the data?
- Distribution of the data (parametric or non-parametric tests).
- What tests are appropriate for the data and are appropriate for the questions or hypotheses that you are testing?
- What conclusions can be drawn from the data and what limitations should be applied when interpreting the data?

- Are there aspects of the work that could be investigated in the future and what design would be appropriate to test these hypotheses?

Testing for differences between groups of data

Once the data have been summarised, you may wish to use a statistical test to help you decide whether to accept or reject your hypothesis. A significant result in data analysis is usually represented as $p < 0.05$, which indicates that you can be 95% confident that your results haven't occurred by chance ($p < 0.001$ indicates a higher degree of significance). At this point either consult a statistician, or use the following information and the key outlined in Appendix B to determine which statistical test to use. Most of the tests outlined can be done in Excel and they are easily done in R if you know how to use it. Another option worth investigating is OpenEpi, which is free to download[3] and provides a robust yet easy-to-use option for simple statistical tests. Note that nowadays, many tests are done by modelling data and most of the tests in the key are not used so much. However, it does not matter which test you do use, provided that the test is appropriate. Also, use the simplest appropriate test—there is no point in using a complex test when a simple one will do.

Comparing two independent sets of data

Some of the more common questions asked in biology relate to whether one set of data significantly differs from another. To answer this, it is necessary to determine whether the mean values are significantly different using a two-sample test of difference.

If the two groups of data are, for example, invertebrate abundance in forest A and invertebrate abundance in forest B, and they are both normally distributed, then an independent *t*-test can be completed or a one-way analysis of variance (ANOVA). If your data are not normally distributed, the non-parametric equivalent is the Mann–Whitney *U* test.

Comparing multiple sets of independent data

One-way ANOVA can also be used if you have multiple groups of data that are independent and normally distributed. The non-parametric equivalent is the Kruskal–Wallis test. Both of these tests will enable you to determine whether the means are significantly different but they will not tell you which groups are different from which (Dytham 1999). Further investigation using the Fisher's least significant difference test will enable you to tease out which groups are significantly different from each other.

If your study is not univariate (has more than one observed variable), e.g. the interaction between invertebrate abundance and factors such as plant richness or temperature, a multivariate ANOVA is required.

[3] http://www.openepi.com/Menu/OE_Menu.htm

Test of difference for paired data

The data may be paired, e.g. invertebrate abundance before and after predator control and the null hypothesis to be tested would therefore be that there is no difference in invertebrate abundance between the two groups. Tests such as the *t*-test are often used but do require the data to be normally distributed (an equivalent non-parametric test would be Wilcoxon's signed ranks test). Non-parametric tests can be a useful check to see if the results concur with the parametric test. A simple test to give an indication of whether there are significant differences between large paired data sets is the sign test. This is essentially done by creating two columns of data and looking to see if there is a consistent pattern of high v. low values. The 'BinomDist' function in Excel will calculate the probability of this occurring.

Tests of difference for repeated measures data

If you have multiple groups of data such as invertebrate abundance monitored over a number years, this can be treated as an extension of paired data and is called repeated measures analysis (Dytham 1999). The Friedman test is the non-parametric equivalent (Dytham 1999).

Testing variables for correlation and association

Correlation

The chi-square test of association can be used on categorical data (qualitative), e.g. where the size of invertebrates was scored as large, medium, small. It is used to determine whether there is a relationship between the categories or not. As with other chi-square tests, it is the difference between the observed and expected frequencies in each of the categories that is important. The test cannot be used with percentages.

Pearson's product-moment correlation (parametric test) or Spearman's rank (non-parametric test) can be applied if you need to test for an association between variables to which a numerical value has been given. For example, a study of grasshopper mate pairs was undertaken to determine whether there is a size association—i.e. do large females attract larger males (Table 2).

Figure 8. *Phaulacridium* grasshoppers copulating on river beds in the Mackenzie Basin.

Table 2. Results from a survey of 10 grasshopper pairs in the Mackenzie Basin to determine whether male and female sizes in copulating pairs were correlated.

Grasshopper pair	Male length (mm)	Female length (mm)
1	6	17
2	5	12
3	7	20
4	8	18
5	5	13
6	6	16
7	6	16
8	5	13
9	7	21
10	5	12

In this example a correlation value of $r = 0.87$ was obtained using the Analysis Toolpak in Excel. This value suggests that there is a strong positive relationship between the sizes of grasshoppers copulating. The $r^2 = 0.76$ suggests that 76% of the variation in the size of grasshoppers of one sex is explained by the size of the other.

Standard linear regression

Standard linear regression determines the strength of a relationship between two variables. Initial examination as well as presentation of these variables can be done with a scatter plot. Linear regression produces a 'best fit' line drawn through a set of points defined by the two variables, e.g. temperature v. distance from the forest edge. Regression also provides measures of the slope of the line and whether it is significantly different from zero. To use regression successfully, you need to be sure that you know which variable is the 'predictor' or 'cause' variable. Regression also assumes that the 'predictor' variable has been measured with minimal variability. In our example, distance from the forest edge is the predictor variable. Regression lines are typically presented with associated 95% confidence intervals and can be calculated using the Analysis Toolpak in Excel and other statistical packages like R. Regression lines are often reported with an r^2 value, which gives an indication of the amount of associated variability.

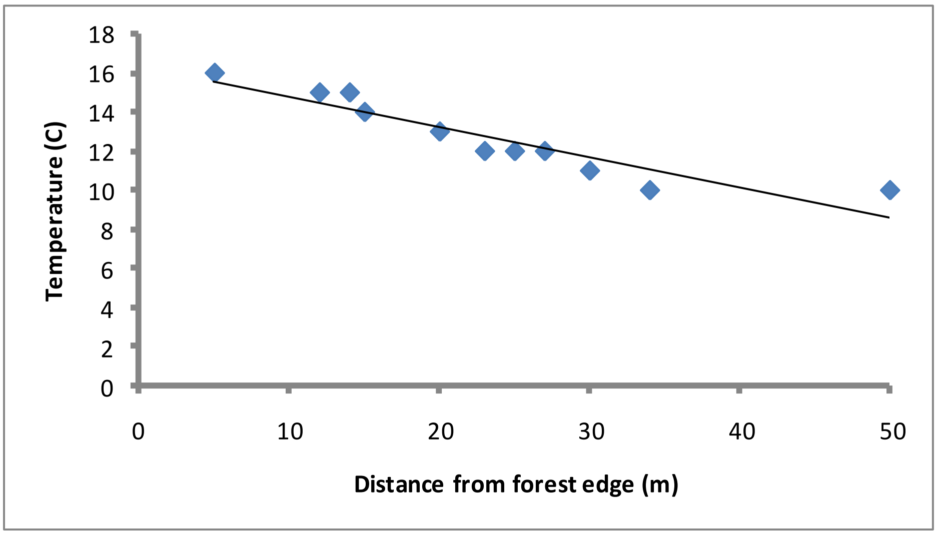

Figure 9: The relationship between temperature recorded at increasing distances from the edge of a beech forest in January 2007 in the Lewis Pass, New Zealand.

A regression line showing the relationship between temperature and distance from the forest edge has been fitted to the data.

In this example (Fig. 9) the r^2 = 0.87 and this means that 87% of the variation in temperature can be explained by the regression line. The output produced by using regression analysis also gave $P < 0.001$, so a highly significant amount of the variation in temperature is explained by distance from the forest edge. The output also confirmed a negative relationship between the variables meaning that temperature decreases as distance from the forest edge increases.

Multiple regression

Multiple regression is an extension of linear regression to more than one ‘predictor’ variable. The assumption with multiple regression is that the ‘predictor’ variables have been measured with minimal variability and that they are clearly independent of each other. Often the ‘predictor’ variables can be confounded or show some correlation and it is recommended to make allowances for this when interpreting the results from a multiple regression. For example, the variables temperature v. distance from forest edge v. density of vegetation may not be considered to be independent. This analysis is not suitable for repeated measures, e.g. monitoring done on annual basis on the same plots. A three-dimensional scatter plot is recommended to display this type of data and any additional ‘predictor’ variables will need to be represented separately.

Choosing which variables to keep or drop from your model

Most modern statistical packages enable a variety of methods for selecting the best subsets of explanatory or predictor variables. Stepwise regression allows more flexibility on which 'predictor' variables are more important (produces lines with slopes that provide a better fit to the data points), and therefore which variables are better at explaining the effects in your study. Guidance on the most appropriate model to select can be found at the following websites:

- http://www.geog.ucsb.edu/~joel/g210_w07/lecture_notes/lect20/oh07_20_2.html
- http://www.palgrave-journals.com/jt/journal/v18/n1/full/jt200926a.html

Consultation with a statistician will help you choose the most suitable method for selection of your final model.

Exploring data for relationships

One approach to conservation of threatened ecosystems is to identify a suite of taxa or indicator species whose diversity can be used to predict the diversity of others (Cranston & Trueman 1997). Because invertebrates are often specific to certain environmental conditions, are linked to ecosystem processes, and are abundant, they make useful indicators. Some species respond to environmental change more than others or have particular attributes that makes them useful for signalling changes in the ecosystem or for monitoring environmental conditions (Hutcheson et al. 1999). The methods used to determine which species are closely related to environmental variables and thus make good indicator species are varied. Multivariate methods (which involve exploring the data using multiple variables) include direct gradient analysis (regression) and indirect gradient analysis (ordination) as well as classification (cluster analysis). Both direct and indirect gradient analysis arrange the sites and/or species data along environmental gradients (Palmer 1993). There are a number of programs available to explore the relationships between communities and environmental factors or determine whether particular groups of species are useful indicators. These include statistical packages such as R, PC-Ord and CANOCO.

Principal component analysis

Principal component analysis (PCA) is similar to regression and shows which variables contribute most to the differences between individuals rather than groups. The method weights all of the variables so that it is possible to discriminate between the individuals that have been assessed. The data have to be normally distributed so the method is not suitable for use with ecological factors. PCA can be used when two or more dimensions have been measured for individual species. For example, the length and width of invertebrates are measured and assessed at varying altitudes on a mountain range. PCA might determine that invertebrate length is the main thing affected by altitude and that the width of the invertebrate is secondary. The output of the analysis shows the weightings that are the strongest, the eigenvalues, which are the degree of relatedness to the variables (or how important the principle axes are), and the position of the individual species along the axis (vertical or horizontal).

Canonical correspondence analysis

Canonical correspondence analysis (CCA) is a method that enables exploration of relationships between invertebrate communities and environmental variables. It is a computer-generated ordination that weights all of the available data against given variables. CCA presents the correlation of species and sites with environmental variables as points in ordination space. It is an indirect gradient analysis. It assumes that a community of species are responding to environmental gradients and can be used to determine to what extent the environmental gradients explain variation in the species data (ter Braak 1986). Neumegan (2006) assessed the invertebrate fauna using pitfall-flight-intercept traps across gradients between native forest and pasture and between native forest and pine plantations. In her study, she used CCA to demonstrate the effect of distance from the forest edge on invertebrate composition.

Figure 10 was generated using the software package CANOCO (ter Braak 1986). These are the results from a study of invertebrates in beech forest, beech forest invaded by Douglas fir, and Douglas fir-only forest at three locations in the South Island. Each number (1–9) represents the invertebrates collected at each of the three sites and in each forest type. Other variables that were measured included the amount of canopy cover, the number of Douglas fir cones, understory plant richness, and associated vegetation cover types (moss, leaf litter, vascular cover, coarse woody debris). This analysis gives an indication of which variables might be influencing the invertebrate community at each of the sites.

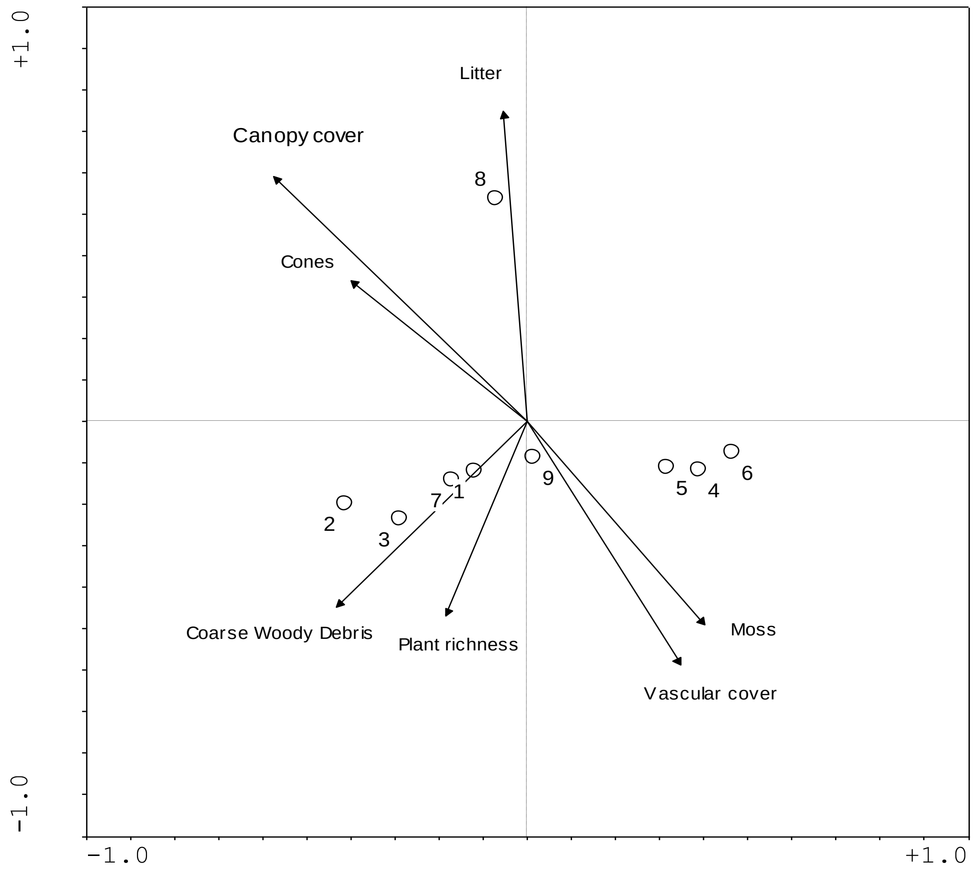

Figure 10. CCA ordination of the invertebrates collected in pitfall traps over a 2-week period at three forests and three forest types (Rankleburn Forest = 1, 2, 3; Burnt Face Forest = 4, 5, 6; Hira Forest = 7, 8, 9) in New Zealand and associated environmental variables.

The location of site scores (1–9) relative to the arrows indicates the environmental characteristics of the sites (Palmer 1993). The environmental variables are presented as arrows and the length of an arrow indicates the strength of the inferred correlation or the importance of the environmental variable (Humphrey et al. 1999). The arrowhead represents the high value for that variable while the medium value is at the centre of the axes. The direction of the arrow shows how the environment is correlated with the abundance of various species or sites and the angle between arrows indicates the correlation between the environmental variables (e.g. in our example moss and vascular cover are correlated). There is also an option in this program to plot all of the individual invertebrate species collected to see which invertebrate species are most ‘influenced’ by the environmental variables. The main limitation of this method is that outliers (atypical data) are not allowed for and can influence the analysis.

TWINSPAN

TWINSPAN (two-way indicator species analysis) is a program also used to explore the relationships between species abundance and environmental variables. The cluster analysis was designed to characterise habitats using indicator species (Hill 1979). It has been used in New Zealand by Hutcheson (1990) and Hutcheson & Jones (1999) to explore whether Malaise-trapped beetles predict the ecological condition of a habitat. The limitations of this method include an assumption that there is a strong relationship between the invertebrate abundance and the environmental variables, and weaker relationships are not considered (Dufrêne & Legendre 1997).

References and further reading

Cranston, P.S.; Trueman, J.W.H. 1997: ‘Indicator’ taxa in invertebrate biodiversity assessment. *Memoirs of the Museum of Victoria 56(2)*: 267–274.

Crawley, M.J. 2002: Statistical computing: an introduction to data analysis using S-Plus. John Wiley & Sons, Chichester.

Dufrêne, M.; Legendre, P. 1997: Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs 67(3)*: 345–366.

Dugdale, J.; Hutcheson, J. 1997: Invertebrate values of kanuka (*Kunzea ericoides*) stands, Gisborne Region, New Zealand. *Science for Conservation 55*. Department of Conservation, Wellington.

Dytham, C. 1999: Choosing and using statistics—a biologist’s guide. Blackwell Science, Oxford.

Eason, C.T.; Gooneratne, R.; Wright, G.R.; Pierce, R.; Frampton, C.M. 1993: The fate of sodium monofluoroacetate (1080) in water, mammals and invertebrates. *Proc. 46th Plant Protection Conference:* 297–301.

Hill, M.O. 1979: TWINSPAN: a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornell University, New York.

Humphrey, J.W.; Hawes, C.; Peace, A.J.; Ferris-Kaan, R.; Jukes, M.R. 1999: Relationships between insect diversity and habitat characteristics in plantation forests. *Forest Ecology and Management 113*: 11–21.

Hutcheson, J. 1990: Characterisation of terrestrial insect communities using Malaise trapped Coleoptera. *Ecological Entomology 15(2)*: 143–151.

Hutcheson, J.; Jones, D. 1999: Spatial variability of insect communities in a homogenous system: measuring biodiversity using Malaise trapped beetles in a *Pinus radiata* plantation in New Zealand. *Forest Ecology and Management 118*: 93–105.

Hutcheson, J.; Walsh, P.; Given, D. 1999: Potential value of indicator species for conservation and management of New Zealand terrestrial communities. *Science for Conservation 109*, Department of Conservation, Wellington.

Magurran, A.E. 1988: Ecological diversity and its measurement. Princeton University Press, Princeton.

McAleece, N. 1997: BioDiversity Pro: Free Statistics Software for Ecology. The Natural History Museum and Scottish Association for Marine Science, London.

Neumegen, R.E. 2006: The influence of adjacent land use on invertebrate communities in native forest fragments on Banks Peninsula, New Zealand. Honours thesis, University of Canterbury, Christchurch.

Palmer, M.W. 1993: Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology 74(8)*: 2215–2230.

Powlesland, R.G.; Stringer, I.A.N.; Hedderley, D.I. 2005: Effects of an aerial 1080 possum poison operation using carrot baits on invertebrates in artificial refuges at Whirinaki Forest Park, 1999–2002. *New Zealand Journal of Ecology 29(2)*: 193–205.

ter Braak, C.J.F. 1986: Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology 67*, 1167–1179.

Appendix A

Example of an Excel spreadsheet suitable for statistical analysis showing the total abundance of four beetle species collected in 10 pitfall-flight-intercept-traps in four different sites.

Site	PF trap	*Megadromus antarcticus*	*Megadromus guerinii*	*Mecodema oregoides*	*Pristoderus bakewellii*
1	1	1	0	3	15
1	2	5	1	6	12
1	3	3	5	2	10
1	4	8	0	6	9
1	5	6	0	5	8
1	6	2	0	5	6
1	7	4	5	2	13
1	8	2	2	4	5
1	9	1	3	2	6
1	10	3	6	4	2
2	1	5	0	2	15
2	2	9	1	6	14
2	3	6	2	3	12
2	4	5	1	3	11
2	5	8	1	3	9
2	6	7	1	3	12
2	7	2	0	3	13
2	8	4	0	2	12
2	9	6	2	1	5
2	10	6	0	2	8
3	1	3	2	1	14
3	2	5	1	2	15
3	3	8	0	1	19
3	4	2	1	0	11
3	5	9	1	0	16
3	6	7	1	5	9
3	7	7	0	2	15
3	8	4	2	3	3
3	9	3	1	1	8
3	10	2	0	3	8
4	1	5	1	3	10
4	2	5	0	2	9
4	3	6	1	5	11
4	4	9	0	2	9
4	5	8	0	4	23
4	6	4	3	2	15
4	7	1	2	2	11
4	8	2	1	2	8
4	9	5	1	0	7
4	10	8	1	1	12

Appendix B

Key to using classical statistical tests (Crawley 2002).

1.	Explanatory variables all categorical	2
	At least one explanatory variable a continuous measurement	4
2.	Response variable a count	*use a contingency table*
	Response variable not a count	3
3.	Response variable a continuous measurement	*use analysis of variance*
	Response variable other than this	*use analysis of deviance*
4.	All explanatory variables continuous	*use regression*
	Response variables both continuous and categorical	*use analysis of covariance*
5.	Response variable continuous	*use regression* or *Ancova*
	Response variable a count	*use log-linear models (Poisson errors)*
	Response variable a proportion	*use logistic model (binomial errors)*
	Response variable a time at death	*use survival analysis*
	Response variable binary	*use binary logistic analysis*
	Response variable is time	*use time series analysis*