

Effect size estimates and confidence intervals: An alternative focus for the presentation and interpretation of ecological data

In A. R. Burk (Ed.) (2005). *New trends in ecology research*. New York: Nova Science Publishers. (pp. 71-102).

© Nova Science Publishers www.novapublishers.com

This document may not exactly replicate the final version published in the book. It is not the copy of record."

Julian Di Stefano

School of Forest and Ecosystem Science and Department of Zoology, University of Melbourne

Fiona Fidler

School of Botany and Department of History and Philosophy of Science, University of Melbourne

Geoff Cumming

School of Psychological Science, La Trobe University

Abstract

Since the development and popularisation of statistical science during the 1930s, the uptake of statistical hypothesis tests in major science disciplines like psychology, medicine and biology has been rapid. Hypothesis tests did not become common in the ecological literature until the 1960s but, since then, testing zero null hypotheses and using P-values to make dichotomous decisions about rejecting them has become routine. Recent surveys of the ecological literature show that P-values are reported in the vast majority of published research articles.

We believe that the focus on tests of zero null hypotheses and the reliance on P-values is inappropriate and argue that ecological data are more relevant to both ecologists and applied users if *a priori* sample size determination procedures are conducted, and effect size estimates and their associated confidence intervals are used to present results. We report evidence that effect sizes, associated uncertainty estimates and sample size calculations are missing from a large number of publications in two leading conservation biology journals. An overwhelming majority (92%) of surveyed articles reported P-values, but only 3% of these reported statistical power. Furthermore, many articles were missing at least one estimate of effect size (43%) or measure of variance (67%). Confidence intervals were reported in only 19% of articles and, when reported, were often inadequately interpreted.

An ecological data set is analysed using both P-values and confidence intervals, and the results from these two approaches compared. We also demonstrate a novel precision-based approach for *a priori* sample size calculations that is consistent with the philosophy of

interval estimation. Although we advocate the use of confidence intervals, recent research shows that they are often incorrectly interpreted. Two major problems are (a) underestimation of the degree to which two confidence intervals must overlap before their associated means are statistically different and (b) overestimation of the proportion of future means enclosed by a sample interval. We provide guidelines for the interpretation and presentation of confidence intervals so that users of ecological data can make the most of the information they contain.

1. Introduction

In this chapter, our primary objective is to present a sensible strategy for the analysis and presentation of ecological data. We believe that, in the majority of cases, the quality and accessibility of research outputs will be improved by (a) conducting *a priori* sample size determination procedures and (b) presenting results using ecologically relevant effect size estimates and associated measures of uncertainty, a process commonly referred to as interval estimation. Proponents of interval estimation are often highly critical of traditional hypothesis testing practices, but we feel it is important to note that interval estimation and hypothesis testing are not mutually exclusive. In fact many ecological studies have as their primary aim comparisons between two or more parameter estimates or between a parameter estimate and an ecologically relevant standard, and as such can be analyses in relation to a null and alternative hypothesis. What we regard as an uninformative practice is testing zero null hypothesis (nulls do not have to be hypotheses of no difference) and the use of P-values to make dichotomous decisions about study results.

Although the analyses we describe operate within a frequentist (or classical)¹ statistical framework, we are not married to a particular data analysis paradigm. In the ecological literature there is growing support for both Bayesian statistics (Ellison 1996, 2004; Wade 2000) and model selection approaches (Anderson *et al.* 2000; Burnham & Anderson 1998,

¹ Throughout this chapter we use the terms ‘frequentist’ and ‘classical’ interchangeably to describe the branch of statistics that draws inferences by comparing a test statistic generated from sample data to the theoretical probability distribution for that statistic. This comparison is used to generate probability statements that are correctly interpreted as a long-run frequency, although many users of frequentist statistical techniques do not interpret the output in this way.

2001), and enhanced graphical presentation of data has also been advocated for some time (Tukey 1977). There are often a number of rational ways to analyse an ecological data set, and Bayesian, model selection or simple graphical analysis are all useful in particular contexts. Our aim is to describe an alternative that is relevant in many situations.

Before beginning our discussion about hypothesis testing, P-values and confidence intervals, we present a brief historical sketch outlining the development of classical statistical inference throughout the first half of the twentieth century, the rapid uptake of null hypothesis testing by a number of disciplines in the years after World War Two, and the more recent debate concerning the utility of this practice. This is presented in section 2. This history has been well documented by others (Fidler *et al.* in press; Gigerenzer 1993; Gigerenzer *et al.* 1989; Senn 1997), and we rely heavily on these sources. Nevertheless, we feel a brief retelling is appropriate for a number of reasons. First, most ecologists are unaware of the historical developments associated with classical statistics and the hypothesis testing debate, and we feel it is an interesting and informative story. More importantly, the history outlines the parallel existence of two different and conflicting analytical approaches and their unreconciled merger after World War Two, and helps to explain many of the problems associated with null hypothesis testing and the interpretation of P-values that exist today. It also reveals the move away from null hypothesis testing by some other disciplines, and strengthens the legitimacy of alternative approaches.

In section 3 we identify null hypothesis testing and the use of P-values as the predominant current framework for the analysis and presentation of ecological data, outline criticisms of this approach, and present survey data that demonstrate the current poor state of statistical reporting in conservation biology journals. In section 4 we describe how effect size

estimation and error presentation using confidence intervals can provide more information than the traditional hypothesis testing approach. Examples are presented that illustrate the use of confidence intervals for both data interpretation and sample size determination. In section 5 we outline some of the challenges confronting interval estimation, and present some preliminary data exploring the way confidence intervals are understood and interpreted by both students and established researchers. We conclude with a call for ecologists to consider alternatives to zero null hypotheses, for research examining cognitive responses to confidence intervals, and for journal editors to promote improved methods for the analysis and interpretation of ecological data.

2. A brief history of classical statistical inference and the hypothesis testing debate

The development of classical statistical science is often attributed to Sir Ronald Fisher who produced a number of influential texts throughout the first half of the 1900s (Fisher 1925, 1935, 1956). Modern classical statistics, however, is really a blend of Fisher's ideas with those of Jerzy Neyman and Egon Pearson (Neyman 1950, 1955, 1957; Neyman & Pearson 1928; Pearson 1962). Neyman and Pearson originally developed their theory to complement and extend Fisher's work, but a series of intellectual and personal disagreements between Fisher on the one hand and Neyman and Pearson on the other resulted in an often acrimonious discourse that lasted from the 1930s until Fisher's death in 1962. Although the debate between the parties ended at this time, the conflict between the theories remains (Gigerenzer 1993).

The inferential framework proposed by Fisher advocated the specification and testing of null hypotheses (Fisher 1935), the modern derivation of which has become known as null

hypothesis significance testing, or NHST. The procedures were based on frequentist principles and involved specifying a null hypothesis and then exposing this null to data. If the data were sufficiently departed from the null, the null was said to be implausible. The degree to which the data and the null had to be separated before the null was considered implausible was defined by Fisher as the ‘level of significance’ and computed (using significance tests such as the *t*-test) as $P(D/H)$, the probability of obtaining the data (or data more extreme) assuming that the null hypothesis was true (such outputs are commonly called P-values, and shall be referred to as such henceforth). This is in stark contrast to Bayes Theorem (Bayes 1763), another prominent inferential technique, which enables the calculation of $P(H/D)$ – the probability of a hypothesis given a particular set of data. This is often what scientists really want to know (Cohen 1994) and can lead to the misinterpretation of P-values, an issue that will be covered in more detail later.

If the P-value resulting from a significance test was low (Fisher suggested 0.05 or 0.01 as conventions), the null hypothesis was considered sufficiently implausible. Fisher considered significant results (ie, P-values below the pre-specified cut off point) to be evidence directly related to the specific null hypothesis being tested (a view that, strictly speaking, is inconsistent with the frequentist precept upon which his inferential framework was based), but believed that non-significant results did not represent evidence one way or the other and should be ignored (Fisher 1935). His framework was asymmetrical in the sense that it only specified a single statistical hypothesis (the null), and only incorporated Type I errors (rejecting the null hypothesis when it is actually true in the real world). Alternative hypotheses, Type II errors (failing to reject the null when it is in fact false) and

statistical power (1 minus the Type II error rate) were rejected by Fisher as inappropriate for scientific inference and discovery (Gigerenzer 1993).²

Like Fisher, Neyman and Pearson believed that inferential procedures should operate within a frequentist framework, but took a more literal approach to the interpretation of frequentist statistical outputs. While Fisher believed that the results of a significance test were directly related to a particular experiment, Neyman-Pearson theory interpreted P-values as the probability of a Type I error over the long term. For Neyman and Pearson, a P-value did not relate to the sample of data used to generate it but to a notional infinite sequence of samples taken from the same population (Neyman 1950).

Neyman-Pearson theory also specified statistical null and alternative hypotheses, one of which was assumed to be true (Neyman & Pearson 1928). This altered the interpretation of non-significant results from 'inconclusive' in Fisher's case to evidence supporting the null hypothesis. The formal specification of two statistical hypotheses also enabled the probability of both Type I and Type II errors to be calculated, and Neyman and Pearson referred to these two values as α (alpha) and β (beta). The formal definition of β meant that the power of a statistical test could be derived, a concept defined as the long run probability of rejecting the null hypothesis (ie, getting a statistically significant result) if it is in fact false in the real world. With two competing hypotheses and a formal definition of both α and β , Neyman-Pearson theory advocated *a priori* cost-benefit analysis to define appropriate Type I and Type II error rates for a particular experimental scenario, and thus

² Fisher's inferential logic was somewhat inconsistent and substantial differences on a number of key points can be found between his earlier and later works. For a detailed account of the development of Fisher's logic and the changes that occurred throughout his life see Gigerenzer (1993) and Gigerenzer *et al.* (1989).

paved the way for modern *a priori* power calculations. Once these *a priori* considerations had been made, however, the alpha level at which to reject the null hypotheses remained fixed and in this sense, Neyman and Pearson were strong proponents of the dichotomous decision to accept or reject the null hypothesis depending on which side of the line the P-value fell. Although this is consistent with Fisher's early work (Fisher 1935), later in his life he changed his mind and criticised this approach as far too restrictive (Fisher 1955); see footnote 2).

There are clearly a number of irreconcilable differences between Fisher's version of frequentist inference and Neyman-Pearson theory. Fisher was not concerned with alternative hypotheses, Type II error rates or sample size calculations, and (at least in his later work) interpreted P-values as evidence for or against a particular hypothesis. Neyman-Pearson theory, on the other hand, specified prior definition of alternative hypotheses and significance levels, and encouraged the calculation of both Type I and Type II error rates. Neyman and Pearson took frequentist inference literally, and interpreted P-values as the long run probability of a Type I error.

In the years immediately following World War Two formalised statistical procedures became increasingly popular in some fields of experimental science, particularly psychology, medicine and associated sub-disciplines (Fidler *et al.* in press; Hubbard & Ryan 2000). This was greatly facilitated by a number of new textbooks written specifically for practitioners that advocated null hypothesis testing. Many of these texts, however, attributed the development of hypothesis testing to Fisher but combined Fisher's ideas with Neyman-Pearson concepts without acknowledging their original source or discussing the different theoretical and somewhat conflicting mathematical and epistemological

frameworks. For example, P-values were often described as evidence for or against a particular hypothesis (Fisher) but Type I and Type II errors explained as a long run probability (Neyman-Pearson). This difference, compounded by the obvious confusion of some of the authors, lead to conflicting messages about the meaning of P-values that worked their way into the consciousness of practising scientists (Gigerenzer 1993).

The net result within the psychological and medical scientific communities was the adoption of a mechanical, apparently unified data analytic framework advocating null hypothesis tests and the use of P-values as the predominant method for presenting results. Although this practice did not become prominent within the ecological community until the late 1960s, since then the uptake of null hypothesis testing has been rapid and adopted as the analytical method of choice in a number of influential books and papers (Downes *et al.* 2002; Quinn & Keough 2002; Sokal & Rohlf 1995; Underwood 1990; Zar 1999). This, combined with the development of easy to use hypothesis testing software and a publication environment biased towards statistically significant results (Palmer 2000), has made null hypothesis tests and P-values the most widely used method of analysing and reporting ecological data.

To be fair, modern statistical texts aimed at the ecological community do warn against some of the problems associated with null hypothesis testing. For example, (Quinn & Keough 2002) include a section outlining the arguments against hypothesis testing, discuss Bayesian analysis and encourage the presentation of effect sizes and measures of uncertainty, while (Sokal & Rohlf 1995) suggest scientists use terms such as 'important' and 'meaningful' to differentiate between statistical significance and biological

importance.³ As the data presented in the next section show, however, these recommendations are often not incorporated into published ecological studies.

Most ecologists are probably unaware of the major debate within psychology, medicine and related fields regarding the use and interpretation of hypothesis tests and P-values. The American Psychological Association, for example, convened a taskforce whose objective was ‘...to elucidate some of the controversial issues surrounding applications of statistics including significance testing and its alternatives...’ (Wilkinson 1999). The report contained many recommendations for the analysis and reporting of data, including a call for the presentation of effect sizes and an associated measure of uncertainty, advice that is repeated in the editorial policies of many medical journals. In some extreme cases (eg, *American Journal of Public Health and Epidemiology*), editors banned the use of P-values altogether (Fidler *et al.* 2004).

The debate is also active among statisticians (Chatfield 1985; Cox 1977; Nelder 1999; Sellke *et al.* 2001). Nelder (1999), for example, refers to the general overemphasis on P-values (the P-value culture, as he calls it) as non-scientific, and contends that the overuse of P-values obstructs the accumulation of scientific information. (Sellke *et al.* 2001) identify that P-values are often incorrectly interpreted and, in tests of precise null hypotheses, a P-value of 0.05 may not provide particularly strong evidence against the null.

³ Although Sokal and Rohlf (1995) differentiate between statistical significance and biological importance, throughout their book they ‘accept null hypotheses’ in response to P-values ≥ 0.05 , an inappropriate and misleading conclusion in many cases.

A number of papers have recently appeared in the ecological literature criticising the way that null hypothesis tests and P-values are used (Anderson *et al.* 2000; Di Stefano 2004; Ellison 1996; Johnson 1999, 2002; Osenberg *et al.* 2002; Yoccoz 1991), and a number of these, and some others, have advocated interval estimation as a complementary or alternative strategy (Anderson *et al.* 2001; Di Stefano 2004; Gerard *et al.* 1998; Johnson 1999; Osenberg *et al.* 2002; Steidl *et al.* 1997; Steidl & Thomas 2001; Yoccoz 1991). We believe these issues deserve serious consideration, and we encourage ecologists to take part in the debate.

3. Statistical inference in ecology

Unlike psychology which embraced modern statistical procedures during the 1950s (Hubbard & Ryan 2000), statistical tests of null hypothesis did not become routine in the ecological literature until the second half of the 1960s (Fidler *et al.* in press). Since then the use of statistical tests, and the reporting of P-values in particular, has increased at a rapid rate. (Anderson *et al.* 2000) surveyed papers published in the highly regarded journal *Ecology* between 1978 and 1997 and found that the average number of P-values per paper rose from 10 - 20 in the late 1970s and early 1980s to 25 - 40 during the late 1980s and 1990s. Remarkably, one of the sampled papers from 1984 contained 317 P-values, and three others from 1991, 1996 and 1997 contained 204, 208 and 208 respectively. (Anderson *et al.* 2000) also surveyed papers from the *Journal of Wildlife Management*, and the results revealed a similar trend. A survey we conducted of 50 papers each from *Conservation Biology* and *Biological Conservation* from 2000 and 2001 issues showed that 92 out of the 100 reported P-values. Together with surveys of journals in other sub-fields (Peterman 1990; Thomas & Juanes 1996), these provide strong evidence that the use of

classical statistical techniques and the presentation of P-values is prevalent in ecological research.

Although we believe that classical statistical procedures are frequently useful for planning experiments and analysing ecological data, there are a number of problems with the way these methods are currently used, interpreted and reported in the ecological literature. These problems have been discussed in detail elsewhere, and the main criticisms are outlined below.

Mechanistic use of statistical procedures. In much of the published ecological literature, statistical procedures are used mechanistically and are frequently viewed as objective mathematical rules that provide unambiguous and rigorous answers to scientific questions. In reality, however, appropriate use of statistical procedures and the interpretation of their outputs involves judgements informed by the theoretical, physical and social context surrounding each study (Stewart-Oaten 1995). The mechanistic use of statistics is a modern phenomenon and was not advocated by Fisher, Neyman or Pearson; a number of references can be found in their work stating the importance of judgement in statistical inference (Gigerenzer 1993).

Testing uninformative null hypotheses. Commonly, statistical null hypotheses state that there is no difference between parameters, or that a parameter equals zero, propositions that are almost always false. It is well documented that the use of statistical tests to reject such nulls is generally trivial and uninformative (Anderson *et al.* 2000; Johnson 1999). As an example, consider an experiment designed to test the general theory that mammals living at sites with high resource quality will have smaller home ranges than the same

species occupying sites with low quality resources. To test this theory, individuals are trapped at both high and low quality sites, and their home ranges recorded. Traditionally, the null hypothesis is specified as $H_0: \mu_1 = \mu_2$, the mean home range size at high quality sites equals the mean home range size at low quality sites. This null is trivial because, regardless of the impact of resource quality on home range size, the mean of one group is almost certain to differ from the mean of the other. It is uninformative because its rejection is inevitable given adequate sample size. Important ecological information, such as the mean home range size of animals living at high and low quality sites, the difference between these means, and the errors associated with these parameter estimates, are not related to the specification of a zero null hypothesis.

Over reliance on P-values. Our survey (presented in detail later), and the survey conducted by (Anderson *et al.* 2000), show that P-values are used frequently in the ecological literature, and often presented as a results summary without additional information such as effect sizes or error estimates.

Uninformative nature of P-values. In many ecological experiments, estimates of effect size and the errors associated with these estimates are the most important outputs from the data (Di Stefano 2004; Johnson 1999). Consider the classical (and very common) form of manipulative experiment where an experimenter imposes different states (treatments) on groups of replicated experimental units and records the effect by measuring a meaningful response variable. In this case, the parameters of interest are (a) the estimated treatment effects, that is, the change in the response variable resulting from each treatment, relative to other treatments or to a control condition, and (b) some measure of how accurate these

estimates are. All P-values do is specify if an effect is likely (or unlikely) to exist. They provide no information about the size of an effect or its precision.

Misinterpretation of P-values. As stated earlier, the P-value is the probability of obtaining data (or data more extreme), assuming that the null hypothesis is true. Put another way, a P-value specifies how consistent (or inconsistent) a data set is with the null hypothesis (Ellison 1996). Nevertheless, many scientists interpret P-values as something different. A number of common misinterpretations are:

- P is probability that observed effects are due to chance.
- $1-P$ is the reliability of a result (ie, the probability of obtaining a similar result if the study were repeated).
- P is the probability that the null hypothesis is true (and therefore $1-P$ is the probability that the alternative hypothesis is true). This is the correct interpretation of a probability statement derived from Bayesian analysis, and is really what most researchers want to know. In fact P-values calculated using classical frequentist approaches can be ‘...highly misleading measures of the evidence provided by the data against the null hypothesis.’ (Berger & Sellke 1987).
- P-values are indicative of effect sizes (ie, small P-values mean large effects).
- P-values are indicative of ecological importance (ie, small P-values mean ecologically important effects). (Yoccoz 1991) provides an eloquent summary of the difference between statistical significance and ecological importance.

Although some ecologists do not interpret P-values incorrectly, many do (Ellison 2004), and these interpretive errors can lead to misinformed and unfounded conclusions.

Use of arbitrary decision criteria. In his book about the use of Analysis of Variance in ecology, (Underwood 1997) has written:

Biologists...have been obsessed with conventional and arbitrary views about the probability of Type I error, so that we use $P = 0.05$ as though it were written on a tablet of stone by some god of statistical theory.

There is no god of statistical theory and a Type I error rate of 5% is not sacred. In fact the notion of specifying a cut-off point below which results are not significant and above which they are is ecologically uninformative (Anderson *et al.* 2000; Johnson 1999). Unfortunately this is still an all too common occurrence in the ecological literature - a browse through a recent issue of *Austral Ecology* at the time of writing revealed the specification of $P = 0.049$ as significant and $P = 0.059$ as not significant (Barrick 2003). As (Rosnow & Rosenthal 1989) have said, 'surely God loves 0.06 nearly as much as 0.05'. The use of an arbitrary cut-off point (whether 0.05 or another value) to specify statistical significance has little ecological meaning and promotes the erroneous notion that every experiment supports an unambiguous conclusion (Smithson 2003).

Under use of a priori sample size determination techniques. Classical statistical techniques can be used in the planning phase of studies to estimate the sample size required to detect an effect of specified magnitude. In the ecological literature this has been most commonly discussed in the context of *a priori* power analysis (Di Stefano 2001; Downes *et al.* 2002; Fairweather 1991; Foster 2001; Keough & Mapstone 1997; Mapstone 1995), but

confidence intervals can also be used. We demonstrate an application of this technique in Section 4.3.

Sample size calculations prior to an experiment are useful as researchers are forced to specify their statistical model and consider the size of ecologically important effects and their expected variances. In the case of *a priori* power analysis, the relative cost of Type I and Type II errors and how these two parameters interact must also be contemplated (Di Stefano 2003; Fairweather 1991; Mapstone 1995). Nevertheless, *a priori* sample size determination is infrequently conducted in ecology resulting in parameter estimates that are often imprecise and uninformative (Fairweather 1991; Mapstone 1995; Peterman 1990).

3.1. Statistical reporting practices in Conservation Biology and Biological Conservation

In order to substantiate claims that statistical reporting practices are often inadequate in the ecological literature, we conducted a detailed survey of 50 *Conservation Biology* and 50 *Biological Conservation* articles published in 2000 and 2001. We surveyed a systematic sample of articles that contained new empirical data, but excluded meta-analyses, methodological and theoretical articles. We recorded a number of variables (outlined below), and calculated the proportion of articles reporting each one, along with the 95% confidence intervals for those proportions using the method described by (Newcombe & Altman 2000). The variables we recorded and the results for each are outlined below:

Null hypothesis tests. Overall, 92% (92 of 100; 95% confidence interval: 85 to 96%) of articles tested statistical null hypothesis.

Null hypothesis of no difference. Of the 92% that tested null hypothesis, 79% (73 of 92; 70 to 86%) tested a null hypothesis of no difference. Due to insufficient information, this may be an underestimate.

Statistical non-significance and statistical power. Most articles (80%, 74 of 92; 71 to 87%) reported at least one statistically non-significant result and about half of these (47%, 35 of 74; 35 to 57%) interpreted the statistically non-significant result as evidence for “no effect” or “no relationship”. However, only 3% (2 of 74; 0 to 9%) reported statistical power. A further 30% (22 of 74; 21 to 41%) made an implicit reference to power issues, such as noting that the sample size was small.

Ambiguous use of ‘significant’. Two thirds of articles (68%, 63 of 92; 58 to 77%) used the word ‘significant’ ambiguously. If the author did not preface ‘significant’ with ‘statistically’, or follow it with a P-value or test statistic, or otherwise differentiate statistical and substantive interpretations, the practice was recorded as ‘ambiguous’.

P-value reporting. One quarter (23 of 92; 17 to 35%) of articles that tested null hypothesis reported statistical significance using asterisks and/or predefined cut-off levels (eg <0.05) while 62% (56 of 92; 51 to 70%) reported exact P-values. Use of asterisks (one, two, or three star significance) has been heavily criticised (Meehl 1978) as this practice provides even less information than exact P-values, is usually insufficient for meta-analysis and has the potential to mislead researchers into thinking that an effect with two stars is more important than an effect with one.

Effect sizes, variance measures, and sample sizes. Most articles reported at least one effect size, variance measure, and sample size, but in many articles at least one of these items was missing. Summary data are presented in Table 1.

Confidence intervals. Of the 100 articles, 19% (13 to 28%) reported confidence intervals, and 26% (5 of 19; 12 to 49%) attempted to interpret them. Any mention of a confidence interval beyond its mere reporting, was recorded as an interpretation.

Figures. A figure was any visual representation of new data and 77% of articles provided at least one. Of these 40% (31 of 77; 30 to 51%) included error bars, representing standard deviation ($n=6$), standard error ($n=14$) or confidence interval ($n=9$). In two cases, bars were unlabelled.

In general, our survey shows that null hypothesis testing and P-values are the dominant analytical and data presentation method used in the *Conservation Biology* and *Biological Conservation*, a result consistent with other surveys of the ecological literature. Specifically, the data indicate that there is a general methodological failure in these journals which requires urgent attention.

Of major concern is (a) the use of P-values alone without associated effect size estimates and measures of precision, (b) the interpretation of non-significant results as evidence of no effect and (c) the lack of *a priori* sample size calculations. For example, about half the articles that reported statistically non-significant results interpreted them as ‘no effect’ or ‘no relationship’, but statistical power calculations were only conducted in two cases. In

addition, effect size estimates and associated measures of precision were not presented for many tests, and were infrequently interpreted when they were reported.

The potential consequences of incomplete statistical reporting, particularly low and unknown statistical power or precision, are serious. In applied ecology and, especially, conservation biology it can result in direct, unanticipated, unacceptable environmental damage. We cannot document specific consequences of the misinterpretations and absences detected in our survey, but possibilities include the failure to act when action was warranted, unnecessary expenditure when action was not warranted, and provision of incorrect advice that affected policy and planning decisions. The severity of potential consequences in the field of conservation biology in particular mean that researchers must provide statistical information in a transparent and accessible way, and be highly assiduous in developing professional standards to remedy reporting deficiencies.

4. An alternative focus for the analysis and presentation of ecological data

In many contexts, interval estimation has considerable advantages over testing zero null hypothesis and reporting results using P-values. The advantages stem from what we feel is the central objective of most ecological studies: to determine if an observed effect (change, impact) is large enough to be important in the context of the ecological system under investigation. Put another way, ecologists are primarily interested in determining if an observed effect is ecologically important, not if it is statistically significant. While P-values provide information about the latter, questions about ecological importance are best answered by estimating the size of effects, and presenting this information with an associated measure of precision. Assuming they are interpreted correctly, there is nothing

inherently wrong with presenting P-values along with effect sizes and their associated errors, but P-values alone cannot be used to infer ecological importance. Our view is that, in many situations, P-values should not be reported as they provide no extra information and act as distractions from questions ecologists are attempting to answer.

In this section, we present an example to show how estimates of effects and associated measures of precision can be used to interpret ecological data. Although precision can be calculated in a number of ways, we prefer confidence intervals because we think they are easier to interpret than other measures, and the remainder of this discussion will relate to them. Our example is based on data presented in (Wade 2000) who generated a hypothetical data set to demonstrate the difference between a classical and a Bayesian analysis. Using these data provided an opportunity to contrast the classical approach described by Wade, an approach using confidence intervals, and a Bayesian analysis. We then use a hypothetical example to show how confidence intervals can be used as a planning tool for ecological studies. First, however, we briefly describe what confidence intervals mean and how they should be interpreted.

4.1. Meaning and interpretation of confidence intervals

Imagine a survey designed to estimate if the abundance of species X living in recently degraded habitat patches (for example, resulting from timber harvesting or mining) had fallen below a critical level. From extensive past sampling it is known that abundance in surrounding undisturbed habitat has a precisely estimated mean of 20 individuals per hectare, and, based on available information about the reproductive strategy of the species, there is concern that local extinction may occur if abundance in degraded patches falls to

about half this level. These considerations guide an *a priori* management decision that remedial action should be taken if abundance falls below 10 individuals per hectare. Initially, a pilot study is established and abundance is measured at five degraded sites. Hypothetical abundance data are 16, 8, 11, 7, and 18 per hectare, and the mean effect size (12.0) and 95% confidence interval (5.98 to 18.02) are shown in Figure 1. Also shown is the P-value function for the confidence interval, a concept we will explain shortly.

Now imagine that this survey was repeated many times, and a 95% confidence interval for the mean calculated on each occasion. Because each sample of data is different, the mean effect and its associated 95% confidence interval will be different on every occasion, but on average, 95% of the intervals will contain the true (unknown) population effect. A common misinterpretation is that we can be 95% sure that an interval from a single sample (eg, 5.98 to 18.02) will contain the population effect, but this is not the case. As with all classical statistical outputs, confidence intervals should be interpreted in the context of notional repeated outcomes.

In many situations, confidence intervals are completely compatible with traditional hypothesis testing procedures - if a 95% confidence interval does not include the value specified by the null hypothesis (usually zero), the null can be rejected at the 5% level (Steidl and Thomas, 2001). In the above example, we know that a test of the null hypothesis that the population effect equals zero will have a P-value less than 0.05 because the confidence interval excludes this null (actually, $P = 0.005$). In an ecological context, however, the null of interest is 10, and a formal test of this null gives a P-value of 0.41. Although the exact value cannot be determined from the confidence interval alone, we know that the P-value will be large as the null value is near the centre of the interval.

If researchers can define an ecologically important effect *a priori*, the information provided by confidence intervals can lead to one of five alternative conclusions that differentiate between statistical significance and ecological importance (Figure 2). In many cases, the interpretation of results using confidence intervals will be more complex than this, and will contain a degree of subjectivity (see Example 1 below). Nevertheless, the use of confidence intervals enables researchers to differentiate between statistical significance and ecological importance and thus interpret their data in an ecologically meaningful way.

So what do confidence intervals from a single data sample tell us about the true population effect? In the above example, the confidence interval can be interpreted as representing a range within which the true effect may plausibly lie (Hoenig & Heisey 2001). Another way of putting this is that, although the estimated effect is 12.0, the data are statistically consistent with effects between 5.98 and 18.02 (Goodman & Berlin 1994), although effects closer to the point estimate are most plausible.

This can be shown with that aid of a P-value function ('witches hat' in Figure 1) which assigns a P-value to every possible value of the population effect (which can be thought of as an infinite array of possible nulls).⁴ Remember that a P-value derived from testing a zero null hypothesis is defined as the probability of observing the sample data (or data more extreme) assuming the null hypothesis is zero. But, as in the example above, a zero null is infrequently of ecological interest. A P-value function simply displays the P-values related

⁴ Although we believe that the P-value function is a useful aid to data interpretation, we do not suggest it should be routinely published. Rather, it should be used behind as an interpretive guide. We show it here, however, as we believe it is unfamiliar to most ecologists.

to every possible null value, from zero through to infinity (Rothman 2002). Another way of thinking about this is that a P-value function shows the compatibility of the sample data with each possible null value. Consider the data in Figure 1. The sample data are clearly most consistent with a population effect of 12 ($P = 1.0$), but not at all consistent with population values below 6 or above 18. Values under the peak of the curve (ie, values close to 12) are all reasonably consistent with the observed data while values under the tails of the curve are less consistent with them. The shape of the curve shows the rate at which values within the confidence interval become less consistent with the sample data. If desired, a P-value can be derived for any null value of interest (in the above example, $P=0.41$ for a null value of 10). As we will show in Section 4.2, this can be used to inform confidence interval based inference.

Confidence intervals also provide information about the repeatability of results. Given an original mean and 95% confidence interval, the average probability that a subsequent mean derived from the same population (a replication mean) will fall within the original confidence interval is about 83% (Cumming *et al.* 2004). This is substantially lower than the 95% that might be expected, and arises due to the presence of two sources of variability, one from the original sample and another from the replication sample. Although 83% is the average probability, this value may be higher or substantially lower for a particular sample of data. Figure 3 shows a simulation of 20 independent samples drawn from the same hypothetical population as the pilot data described above. If the mean of a particular sample happens to fall very close to the population value (for example, the third sample mean from the left), and that sample has an interval of typical width, the confidence interval will capture around 95% of replication means. For sample means that happen to fall a little distance away from the population value, however, the confidence

interval is likely to capture distinctly fewer. In the extreme case of a confidence interval that does not capture the population effect (5% of 95% intervals; see the open square cases in Figure 3), less than 50% of future replication effects will be captured.

It is somewhat frustrating that confidence intervals do not provides us with the probability that the interval contains the true effect, a value that would be particularly useful – to achieve this we would have to create intervals using a Bayesian approach. Within the classical statistical framework the best we can do is to say that a 95% confidence interval is very likely but not certain to contain the true effect.

4.2. Example 1: Using confidence intervals to interpret ecological data

(Wade 2000) generated data sets for two hypothetical animal populations (population 1 and population 2; Figure 4) to explore the differences between a classical hypothesis testing approach and a Bayesian analysis. Wade presented graphs of the data including the line of best fit, the regression equations (Figure 4 caption) and summarised the results of the hypothesis tests (null hypothesis of zero slope) using P-values ($P = 0.048$ for population 1 and $P = 0.053$ for population 2). In addition to Wade's data presentation and analysis we have drawn 95% confidence intervals around the line of best fit (dashed lines in Figure 4) and calculated 95% confidence intervals around the slopes; the slope (95% CI) for population 1 is -0.36% (-0.004 to -0.72%) and for population 2 is -10.0% (0.19 to -20%). In addition, we have calculated a P-value function for these intervals to aid interpretation (Figure 5).

In a conservation context, data like those in Wade's hypothetical example are collected to detect if a population is declining at an ecologically important rate. For the purposes of his example, (Wade 2000) assumed that a decline of 5% per year would be ecologically important, and we will adopt the same value here.

In his exposition of the classical hypothesis testing approach, (Wade 2000) assumed that management intervention to halt a decline would only occur if the gradient of the regression line was statistically less than zero. Operating within this framework, population 1 is assumed to be declining ($P = 0.048$) and deserving of management action, while population 2 is not ($P = 0.053$). As Wade has said, this conclusion is irreconcilable with the data. Simply looking at the graphs (Figure 4) clearly shows that population 2 may well be declining at a much greater rate than population 1.

This example raises two further issues regarding the dichotomous decision-making approach described above. First, it demonstrates the nonsensical nature of an arbitrary cut-off point (in this case a P-value of 0.05) for decision-making purposes. Because both P-values are so close to 0.05, a small change in either data set could easily result in a different decision, even though the overall trends would be substantially the same. We would not use P-values to interpret these data, but if presented with nothing else we would conclude that the two data sets provide approximately equal statistical evidence for a population decline. Second, the data from population 1 are precisely estimated and indicate a small decline while the data from population 2 provide a much less precise estimate and indicate a decline that may be small or very large. The calculation of the P-value requires both these pieces of information (the precision and the effect size) and so their independent influence is hidden. As this example shows, the combination of small effect and high

precision in population 1 and large effect and low precision in population 2 produce approximately equal P-values, even though the data from the two populations are of a very different form. The estimated effect and the precision of this estimate are important in their own right and should both be reported.

(Wade 2000) dismissed the hypothesis testing approach described above as inappropriate and advocated a Bayesian solution to the problem. Using an uninformative prior distribution, Wade's Bayesian analysis indicated that the probability of a decline greater than 5% per year is 0 for population 1 and 0.86 for population 2. In addition, the probability that the decline is between 0 and 5% is 0.98 for population 1 and 0.12 for population 2. These data suggest that population 1 is very likely to be declining, but at a rate less than 5% per year while population 2 is probably declining at a rate of greater than 5% per year. We agree that this is a sensible way to analyse the data and the conclusion would be an appropriate guide for management action.

Interval estimation, however, can also be used to interpret these data in a rational way. The statistical technique is classical, but the focus is on effect size estimates (the regression slopes) and the associated confidence intervals.

In percentage terms, population 1 is estimated to be declining at 0.36% per year, but the 95% confidence interval indicates that the data are statistically compatible with declines between 0.004% and 0.72% per year. Because the confidence interval does not include a decline of 5%, we conclude that the data are not statistically compatible with a decline of this magnitude (Case 3 in Figure 2). This is clearly demonstrated by the P-value function (Figure 5) which shows that the data are most consistent with a small decline, but

completely inconsistent with a decline greater than about 1% per year ($P=0$ for declines greater than this value). In contrast, population 2 is estimated to be declining at 10% per year, but the data are statistically compatible with a slight population increase (0.19%) or a decline of up to 20% per year. Population 2 is almost certainly declining, but because the data are variable, we cannot predict the true rate of decline with precision. We feel there are three possible ways to deal with this kind of situation.

- (a) Acknowledge that the population is almost certainly declining, and may well be declining at $>5\%$ per year, but conclude that more data are needed before a decision can be made regarding management intervention. We might, for example, specify that we require enough data for the upper bound of the 95% confidence interval (the bound closer to zero) to exclude a 5% per year decline (Case 1 in Figure 2) before recommending management action. This approach is conservative and weighted towards a conclusion that a critical decline is not occurring.

- (b) Accept that the effect size estimate is imprecise, but recommend management intervention because an ecologically important decline may well be occurring. This conclusion is based on the fact that (a) the best estimate of the decline is 10% per year, and (b) the ecologically important decline (5% per year) is in the right hand tail of the P-value function (Figure 5) making larger (more negative) declines much more probable than smaller declines. As illustrated in Figure 5, the P-value associated with a decline of 5% per year is 0.29, and increases rapidly for larger values. We interpret this to mean that while declines less rapid than 5% per year are plausible values for the true (population) decline, declines more rapid than 5% per year (declines in the range of 8 to 12%, for example) are much more consistent with the sample data.

(c) Accept a higher level of error and use a lower confidence level (eg, 90% or 80%).

From the P-value function it can be seen that a 71% confidence interval would just exclude the ecologically important effect. There is strong evidence that in many conservation scenarios it is more sensible to weight the decision process in favour of effect detection (Fairweather 1991; Peterman 1990; Taylor & Gerrodette 1993), so using a lower confidence level may well be justified. Having said this, we believe that the confidence level should be kept as high as possible, and, if it is lowered, this should be done during the planning phase of a study. We provide some additional comments in Section 4.3.

We feel that option (b) above is the best approach. The confidence interval clearly shows that the point estimate lacks precision and the sample data are statistically consistent with a wide range of possible population declines. The P-value function, however, demonstrates the consistency of various population declines with the sample data, and indicates that declines between zero and 5% per year are much less consistent with the data than a range of declines greater (ie, more negative) than 5% per year. Thus our conclusion matches that of Wade's Bayesian analysis; population 2 is probably declining at a rate greater than 5% per year, and management intervention should occur.

It is important to point out that the inferential process we have just described contains two decisions based on judgement.⁵ The first is the use of a 95% confidence interval (as

⁵ In the context of a real study, many judgements concerning the experimental design, data collection protocols, measurement variables, etc, are made prior to analysis. All of these decisions have the capacity to influence the results.

opposed to a 90% or 80% interval), and the second is the way we interpreted the P-value function as evidence for a >5% decline. Reasoned subjectivity is a necessary part of all data analytic processes whether we like it or not (Stewart-Oaten 1995). The Bayesian approach described by (Wade 2000) also contained judgements. Wade, for example, decided to use an uninformative prior distribution, but it would have been quite acceptable to use an informative prior, a decision that would have altered the output of the analysis. In addition, Wade describes the probability of population 2 declining at >5% (0.86) as ‘fairly high’, and indicates that the Bayes factor comparing the hypotheses ‘slope = zero’ to ‘slope = -5%’ provides ‘positive’ evidence of a 5% decline. Wade concludes (and we agree) that, on the basis of the available evidence, population 2 is probably declining at a rate of at least 5% per year, but this is a reasoned judgement and not an indisputable fact.

In reality, ecological data are highly variable, and, combined with the small sample sizes associated with many ecological research projects, effect size estimates are frequently surrounded by large error margins. Using confidence intervals to interpret ecological data enables this error to be quantified and displayed in a form that is readily appreciated, and facilitates its transparent incorporation into the decision-making process.

4.3. Example 2: Using confidence intervals to plan ecological studies

It is widely acknowledged that power analysis is a useful tool in the planning phase of ecological studies (Di Stefano 2001, 2003; Downes *et al.* 2002; Fairweather 1991; Foster 2001; Gerrodette 1993; Keough & Mapstone 1997; Quinn & Keough 2002; Steidl & Thomas 2001), but the utility of confidence intervals for study planning has rarely been discussed in the ecological literature (but see (Steidl & Thomas 2001)). Nevertheless,

confidence intervals have been considered as a planning tool in other disciplines (Goodman & Berlin 1994; Smithson 2003) and we feel they are just as applicable to ecology.

The use of confidence intervals for study planning has been discussed within both hypothesis testing (Daly 2000; Goodman & Berlin 1994; Smithson 2003; Steidl & Thomas 2001) and interval estimation (Algina *et al.* 2002; Algina & Olejnik 2000) frameworks. The former involves determining sample size using statistical power analysis and then substituting this value into a confidence interval equation to examine the corresponding expected confidence interval width. We refer to this as the power approach. The alternative is simply to use a confidence interval equation to calculate the sample size needed to generate an interval of specified width. We refer to this as the precision approach. As has been noted elsewhere (Algina *et al.* 2002; Algina & Olejnik 2000; Smithson 2003) these methods often produce quite different answers, because the sample size necessary to test a hypothesis with adequate power and to estimate an effect with adequate precision may well be different.

There is in addition a more fundamental distinction between the two approaches. Power analysis is inextricably linked to statistical significance testing and requires the specification of point null and alternative hypotheses. The difference between these hypotheses is the ‘important effect size’ integral to all power calculations, and outputs of this process can only be interpreted in relation to the effect size value used in the calculation. In contrast, the precision approach requires a specification of the desired precision (defined as w , the confidence interval half width, also known as the *margin of error*) with which to estimate the parameter of interest. Consequently we can use this

method to calculate the sample size required to estimate a parameter with a specified degree of precision regardless of its true population value. If we were using the power approach, a value for the true population parameter would need to be specified.

Although the simplicity of the precision approach seems appealing, it has been criticised for resulting in sample sizes that may be considered too small, as the probability that the interval width actually observed in a future study will be greater (or less) than the planned width is approximately 50% (Daly 2000; Goodman & Berlin 1994). In many research scenarios, only having a 50% chance of obtaining one's specified level of precision may be unacceptable. Due to this problem, most of the literature on this topic suggests using power analysis for at least some components of the study planning process, even if confidence intervals will be used to interpret the results. However, we feel the use of statistical power analysis to determine sample size is inconsistent with interval estimation philosophy, and a modification of the precision approach is required.

Here we suggest a novel precision-based method for determining sample size that provides a solution to the problem noted above. For the purpose of comparison, we demonstrate the process using both statistical power analysis and our novel precision method with reference to the hypothetical example outlined in Section 4.1.

The power approach. Let us return to the example outlined in Section 4.1 (Figure 1). Recall that the objective was to determine if abundance at degraded sites had fallen below a critical level which, on the basis past data and informed opinion, had been set at 10 individuals per hectare. As the data from the pilot study were inconclusive, we may wish to

know how many degraded sites need to be surveyed to be confident that abundance has or has not fallen below the critical level.

To perform the power calculation we need to specify an important effect size, values for α and power, and use the standard deviation of the pilot data as an estimate of the population standard deviation. In this case, the important effect size is the smallest difference between the future sample mean and our critical value that we think it is important to detect. We might choose this value to be 2 as we are prepared to accept that abundance estimates between 8 and 12 (within 20% of the critical value) are functionally equivalent. Consequently, the null hypothesis is that abundance at degraded sites is equal to 10, while the alternative is that abundance is 2 units away from 10, in either direction. We are interested in this two tailed alternative as, for management purposes, it is important to determine whether abundance is less than or greater than the critical level.

Rather than deciding on specific values of α and β we follow the advice of (Keough & Mapstone 1997; Mapstone 1995) and others and determine a relevant $\alpha:\beta$ ratio by considering the relative cost of Type I and Type II errors. From a management perspective, making a Type I error would mean investing resources to increase abundance in degraded patches when such an investment was not necessary, while making a Type II error would mean there would be no management action even though it was warranted, resulting in the possible local extinction of the species under consideration. For the purpose of this example let us assume that after thorough discussion with all relevant stakeholders, it is decided that Type I and Type II errors are equally costly, thus defining the $\alpha:\beta$ ratio as 1:1.

Because specific levels of α and power are not defined, we start with a high value for power. Entering power = 0.95 and $\alpha = 0.05$ into a suitable power calculator (Lenth 2000), the number of degraded sites that need to be surveyed to detect an effect size of 2 is 78. Inserting $n = 78$ into the appropriate confidence interval formula gives an expected 95% confidence interval half width of 1.09, although the probability of observing a value of w (the confidence interval half width) no greater than this in future data is only about 50%. If there were not 78 degraded sites or the money was not available for such extensive sampling effort, we would reduce power and increase α so that the 1:1 $\alpha:\beta$ ratio is maintained. For example, when power = 0.90 and $\alpha = 0.10$, 52 degraded sites are required, and the corresponding 95% confidence interval half width is 1.35, and when power = 0.80 and $\alpha = 0.20$, we need 27 sites and the interval half width is 1.92. Until now, this kind of process has been the best way to explore the relationship between sample size, statistical power and confidence interval width.

The precision approach. If the half width of a confidence interval is w , $w = t_{(df),C} \times SE$, where t is the two tailed critical value from a t -distribution with df degrees of freedom, C is the confidence level and SE is the standard error. Further, we differentiate between w_{exp} and w_{obs} , the expected (as calculated pre-study) and observed (post-study) values of w respectively. We use the fact that w is calculated from the sample standard deviation whose sampling distribution is related to the chi-square distribution. We can therefore calculate the one tailed w_{upper} value for any chosen probability PP so that w_{obs} will be less than w_{upper} , for a specified C and sample size. We use PP for the *precision probability*, the probability that our observed precision (w_{obs}), which varies from sample to sample, will be no greater than w_{upper} . PP can also be thought of as the chance that the observed confidence interval will be at least as precise as planned. This process provides a calculation based on

the trade-off between n and w_{upper} , for our chosen PP . No null or alternative hypotheses need be specified, but if we do wish to consider precision in relation to some chosen effect size, we could choose w_{upper} equal to that effect, then calculate the corresponding sample size.

For our calculations, we might specify our desired precision, w , to be 2, for similar reasons as outlined in the power example. We then might want to know the sample size required to have a $PP\%$ chance that the observed precision, w_{obs} , will be less than 2 when the confidence level, C , is set at 95%. Using the precision approach, we find that 25 degraded sites are required to have a $PP = 50\%$ chance of obtaining w_{obs} less than 2 in a future sample. For $PP = 80\%$ and 95% , the required sample size is 30 and 35 respectively. If we wished, we could specify different levels of precision and investigate the affect on sample size. For example, if $w = 1$ and $C = 95\%$, the sample sizes required to achieve a PP of 50%, 80% and 95% are 93, 104 and 114 respectively. Using the standard deviation estimate from the pilot data, Figure 6 shows the relationship between sample size and PP for $w = 1, 2$ and 3 . C is 95% in all cases.

In summary, our novel precision approach to study planning enables researchers to vary the precision probability, PP , and thus examine the relationship between this value, sample size and a specified degree of precision. We believe this method for sample size determination is philosophically consistent with the interval estimation approach advocated in this chapter, and by enabling researchers to vary the precision probability we provide a solution to the problem of inadequate sample size that has been the basis of criticism in the past. If they wish, researchers can also vary the confidence level, C , and observe how this affects the outcome. If possible, however, C should remain high, as the lower C becomes

the less confident we are that values within the confidence interval contain the population parameter. Although we recognise that smaller values of C may be relevant in some situations, we encourage researchers to keep the confidence level as large as possible.

5. Using confidence intervals wisely: some identified misconceptions and guidelines for interpretation

In the previous section we described some advantages associated with a confidence interval approach to interpreting ecological data. However, confidence intervals are not a panacea to all statistical reporting problems; they too are misinterpreted, and the extent to which they are able to supplement or replace P-values is not clear. Although research in cognitive psychology has explored many of the misconceptions associated with P-values and tests of zero null hypothesis (Oakes 1986; Tversky & Kahneman 1971), corresponding work on confidence intervals is relatively undeveloped. Claims regarding the benefits of confidence intervals include the relative ease with which they are understood and taught, their capacity to facilitate better communication of research results and their infrequent misinterpretation (Schmidt 1996; Schmidt & Hunter 1997). Others admit that confidence intervals may be misinterpreted, but that the consequences of such misinterpretations are likely to be less serious than corresponding errors associated with P-values (Hoenig & Heisey 2001). Nevertheless, these statements contain no supporting evidence and the empirical questions they raise are largely unanswered. It is clear that additional research in this area is required.

5.1. Avoiding established misconceptions: 'statistically non-significant equals no effect'

One of the misconceptions associated with P-values (Section 3) is that statistically non-significant results (usually a P-value >0.05) are often equated with the absence of an ecologically important effect. To test the hypothesis that the interpretation of results using confidence intervals reduces the occurrence of this error, we presented 79 final year or postgraduate ecology students with two similar research scenarios that reported results using P-values and confidence intervals respectively. The results of both scenarios were statistically non-significant, but there was a clear specification that statistical power was low and that the observed effect sizes were non trivial. Using a five point Likert scale, students were asked to interpret the data as evidence for or against the null hypothesis of no effect (Fidler *et al.*, unpublished data).

When statistically non-significant results were presented using P-values, almost 40% (31 of 79) of students believed that the data provided positive evidence for the null hypothesis of no effect (the wrong answer), despite the presence of statistical power information and additional guidance about the ecological importance of observed effects. However, the overwhelming majority of these students (87%, 27 of 31) changed their mind when they viewed similar results presented using confidence intervals. Interestingly, of the 60% or so that interpreted results correctly when presented with P-values, 17% (8 of 48) misinterpreted the results when they were expressed using confidence intervals. Thus although confidence intervals usually facilitated the correct interpretation of results, they occasionally resulted in the reverse effect.

To test whether this reverse effect was an artefact of different scenario content, the study was repeated so that each student received the same scenario presented once with P-values and once using confidence intervals. This time we tested 55 second year ecology students,

asking the same question. The results of this study were almost identical to the first except that the reverse effect was reduced by about half. An additional result was the apparent presence of a learning effect. Students who saw the confidence intervals first gave the correct answer on the P-value presentation more often than students who saw P-values first, but there was no corresponding beneficial transfer of seeing the P-values before confidence intervals.

Relative to research scenarios where results were presented using P-values, the students we tested were less likely to interpret statistically non-significant results as evidence for ‘no effect’ when confidence intervals were used. Based on these preliminary data, confidence intervals appear to assist the precautionary interpretation of low powered, statistically non-significant results.

5.2. Creating new misconceptions?

In other surveys we have found confidence intervals susceptible to unique misconceptions not commonly associated with P-values. In two separate surveys of undergraduate students in psychology and ecology (total $n=357$), we found that many failed to recognise confidence intervals as inferential (Fidler & Finch, unpublished data). In a variety of contexts and question types most students interpreted confidence intervals as relating to the data sample from which they were derived, but not to the relevant population parameter. In students’ own definitions, they described the confidence interval as a range or a truncated range of sample observations or measures, rather than as a plausible set of values within which the population parameter is likely to lie.

It is not surprising that undergraduate students have a developing understanding of statistics and thus misinterpret confidence intervals to some degree. Recent research, however, shows that misconceptions about confidence intervals, along with standard error bars, extends to experienced researchers in a number of disciplines, including medicine where confidence intervals are routinely used. Three of these interpretive errors are outlined below.

The overlap misconception. Many researchers falsely believe that for two independent group means to be statistically significantly different, the 95% confidence intervals around those means must not overlap (Schenker & Gentleman 2001). In fact, 95% confidence intervals can overlap by roughly a quarter and the mean difference still be significant at the 5% level (Figure 7). In an internet-based study, (Belia *et al.* submitted) tested the extent of the overlap misconception by inviting authors published in leading psychology, medicine and behavioural neuroscience journals to adjust a figure until they judged two means, each with 95% confidence intervals, to be just statistically significantly different ($P = 0.05$). Only 17% (24 of 140) of respondents adjusted the figure so that corresponding P-value was close to 0.05. Most respondents were too conservative, setting the confidence intervals to just touch or not overlap at all. The mean response positioned the confidence intervals at the equivalent of $P = 0.009$.

Drawing from the same population of researchers, (Belia *et al.* submitted) conducted the same experiment when the means were surrounded by standard error bars (roughly half the width of 95% confidence intervals). In this case, 25% (44 of 179) of respondents adjusted the figure so that corresponding P-value was close to 0.05. As with the confidence interval based exercise, most researchers set the means so that arms of the standard error bars

almost touched, which in this case led to answers that were too lax; the mean response corresponded to $P = 0.109$. In fact, standard error bars must be separated by about one standard error (ie, one arm of a standard error bar) for the corresponding P-value to be about 0.05 (Figure 7). Further discussion concerning the interpretation of error bar overlap can be found in (Cumming & Finch submitted; Saville 2003; Wolfe & Hanley 2002).

Error bar interpretation and experimental design. Figure 8 is an example of a graph displaying typical repeated measures data, including control and treatment means, with associated error bars, at a number of different times. Unknown to many researchers is that error bars around individual means are useful only for interpreting data from the same time interval. Only error bars around the *difference between times* allow us to look directly at the effect of time, which is usually of interest in studies containing repeated temporal measurements. With reference to Figure 8, error bars can be used to assess the difference between Control and Treatment means at Time 1, (for example, by using the overlap guidelines in Figure 7), but the error bars are irrelevant, even misleading, for comparing, for example, Treatment means at Time 1 and Time 2.

Again drawing from a population of published authors in psychology, medicine and behavioural neuroscience, (Belia *et al.* submitted) tested the extent to which this issue was understood. Only 11% (18 of 159) of respondents realised that bars around individual means could not be used to determine statistical significance in repeated measures cases. It may well be that current graphical conventions promote mistakes of this kind, as adjacent error bars like those displayed in Figure 8 invite invalid comparison between correlated data, and traditional figure design does not make salient whether an independent variable is a repeated measure or not.

The confidence level misconception. This is the mistaken belief that a 95% confidence interval will, on average, capture 95% of effects from future samples (replication effects). As we mentioned in Section 4.1, however, a 95% confidence interval only captures on average about 83% of replication effects. In a second internet-based study, (Cumming *et al.* 2004) contacted researchers and asked them to estimate where 10 replication means might plausibly fall given an observed mean and its 95% confidence interval. Almost 80% (105 of 134) of respondents placed 9 or 10 of the replication means within the confidence interval. In open-ended comments many made clear that they had attempted to match the number of future means within the intervals to the confidence level (ie, 95%). Corresponding results were found in a group asked to perform the same task with standard error bars. Most respondents placed 6 or 7 of the replication means within the standard error limits, again matching the number of means to the confidence interval (68% for standard errors). These results indicate that researchers often have a poor understanding of the information error bars provide about where replication means are likely to fall.

5.4. Guidelines for interpreting confidence intervals

Studies conducted to date suggests that a number of misconceptions and interpretive problems may occur when confidence are used to interpret research results. Although these studies have investigated responses of scientists in psychology, medicine and related fields, we expect the same interpretive errors would be made by ecologists. Despite this, we encourage the use of confidence intervals as we believe they are superior to tests of zero null hypotheses and their associated P-values. To support the increased use of confidence intervals we present a number of guidelines designed to minimise interpretive errors.

Further guidance for interpreting confidence intervals can be found in (Cumming & Finch submitted).

- Consider each value within a 95% confidence interval as a plausible estimate of the population parameter, but, although it is probable, there is no guarantee that the population parameter lies within the interval.
- Note that the population parameter is more likely to be near the centre of a confidence interval, near the point estimate, and less likely to be near the outer limits.
- Consider whether a confidence interval contains an ecologically important effect. If so, look at where it is positioned. As mentioned above, effects closer to the centre of the interval are more plausible.
- For a comparison of two independent means, $P \approx 0.05$ when the arms of the 95% confidence intervals overlap by about half the length of one arm, that is, by about one quarter of the total confidence interval width. $P \approx 0.01$ when the overlap is zero or there is a gap between the two intervals. (This applies when both sample sizes are at least 10, and the two interval widths are not substantially different; see Figure 7).
- In studies that include a repeated measure, such as those in which measurements are taken at several times, interval estimates (whether confidence intervals or SE bars) on means at individual times may not be used to guide inferences that involve the repeated measure. See Figure 8.

- Note the width of the interval and what this tells you about the precision of the study. A wide interval indicates that the population parameter has been estimated imprecisely. In general, studies with wide confidence intervals have low statistical power to detect effect sizes likely to be of interest.

6 Conclusions

Based on surveys of the literature it is clear that tests of zero null hypotheses and the use of P-values to make dichotomous decisions about the significance or otherwise of results is widespread in ecology. However, we feel there is good reason for ecologists to consider alternative approaches. The approach outlined here, interval estimation, involves estimating effects and presenting them along with their associated confidence intervals. When confidence intervals are reported, time should be taken to interpret and discuss the information they convey. Wherever possible, ecologists should determine ecologically relevant effect sizes prior to data collection and use these values to aid sample size determination. We believe that interval estimation is highly compatible with the analytical objectives of most ecological studies and thus facilitates a transfer of accurate information to both investigators and applied users.

Proponents of interval estimation often assume that this technique is free of the problems associated with hypothesis testing. This assumption, however, has not been rigorously tested. The preliminary data presented in this chapter indicate that confidence convey more meaning than P-values, but a number of misconceptions still occur. In many instances, the cognitive processes associated with the interpretation of confidence intervals are unclear. In addition, there is some debate regarding the best way to use confidence intervals for

study planning, and we contribute to this by suggesting a novel strategy based on precision. Clearly, a greater degree of consideration and research into these aspects of interval estimation is required.

We believe that journal editors have a major role to play in encouraging the use of interval estimation and other sensible data analysis strategies. Editorial policy outlining the benefits of *a priori* sample size calculations, the specification of ecologically important effect sizes and the use of confidence intervals to report the error associated with point estimates would send a strong message to potential authors that these practices are preferred. While some ecology journals have such guidelines (*Ecology*, and the other Ecological Society of America publications, for example), most do not. As (Ellison 2004) has shown, however, the presence of guidelines do not solve the problem of misinterpreting and misreporting P-values, and comments on manuscripts by editors and reviewers should alert authors to these problems. In one sense, editors are the gatekeepers of ecological data, and thus have tremendous potential to improve statistical reporting practices across the discipline.

Acknowledgments

Russ Lenth, Graham Hepworth, Paul Wade, Lauren Bennett, Mark Burgman, Kylie McKenzie, This work was supported by the Australian Research Council.

Table 1. Summary of reporting practices in a sample of 100 empirical articles from *Biological Conservation* and *Conservation Biology* with respect to effect size, standard deviation, standard error, sample size and confidence intervals.

Reporting practice	% of articles	95% CI
Reporting at least one effect size	87 (80 of 92)	80-94
Missing at least one effect size	43 (40 of 92)	34-53
Reporting at least one SD or SE	48 (44 of 92)	38-58
Missing at least one SD or SE	67 (62 of 92)	58-77
Reporting at least one sample size or df	76 (70 of 92)	68-84
Missing at least one sample size or df	36 (33 of 92)	26-45
Reporting at least one CI	19 (19 of 100)	11-27
Interpreting CIs	26 (5 of 19)	3-39

Figure 1. Estimated mean effect (12.0) and 95% confidence interval (5.98 to 18.02) for hypothetical abundance data. The solid line is the P-value function which specifies the consistency of the sample data with each possible abundance value. The vertical dashed line indicates that the data are most consistent with an abundance of 12. Note that P-values (left Y-axis) are directly related to the confidence level (right Y-axis), as is illustrated by the lower and upper limits of the 95% confidence interval cutting the P-value function at 0.05.

Figure 2. Interpretation of results using confidence intervals. Black squares are estimated effects, error bars are 95% confidence intervals and the dashed line represents an ecologically important effect. In Case 1, the observed effect is both statistically significant and ecologically important. In Case 2, the effect is statistically significant, but the data are insufficient to determine ecological importance. In Case 3, the effect is statistically significant but not ecologically important. In Case 4, the effect is not statistically significant and the data are insufficient to determine ecological importance. In Case 5, the effect is neither statistically significant nor ecologically important. After (Fox 2001) and (Steidl & Thomas 2001).

Figure 3. Results of 20 independent samples drawn from the same population as the pilot data. For the purposes of this illustration we assume that population to have mean = 10, standard deviation = 4, and we take samples of size 5. The horizontal line is the population mean, which here is 10 but in practice is never known. Means of successive samples are shown as squares, each with a 95% confidence interval. The first sample is our pilot data. Running a single experiment is equivalent to choosing a single mean and 95% confidence interval randomly from an infinitely large set of means and confidence intervals, just 20 of

which are shown here. Note that most but not all confidence intervals capture the population mean; in the long run, 95% of the intervals would capture it. Here, two confidence intervals do not include the population mean, and those sample means are shown as open squares. The figure illustrates that the population mean is more often captured near the centre of a confidence interval than near its extremities. Note that these intervals vary considerably in width because the sample size in our hypothetical example is small.

Figure 4. Population 1 and population 2 from (Wade 2000). The solid line is the line of best fit and the dashed lines are 95% confidence intervals (data have been back transformed for easier interpretation). The main graphs show the data at the same scale for the purpose of comparison, but population 1 data are enlarged in the inset. Population 1: $\log_e y = 4.62 - 0.0036x$; 95% CI for the slope = -0.00004 to -0.007; $P = 0.048$; Population 2: $\log_e y = 5.01 - 0.10x$; 95% CI for the slope = 0.002 to -0.20; $P = 0.053$.

Figure 5. P-value functions (solid lines), estimated population declines (black squares) and associated 95% confidence intervals (error bars) for populations 1 and 2. The inset shows data for population 1 on a readable scale. For population 2, the dotted lines show the P-value and confidence level corresponding to a decline of 5% per year. The P-value is 0.29 and the confidence level is 71%.

Figure 6. The relationship between sample size and the precision probability (PP ; the chance that the confidence interval half width observed in a future sample will be less than a specified value, w) for three different values of w . In this case, the confidence level (C) = 95% and standard deviation = 4.85.

Figure 7. Guidelines for interpreting the difference between the means of two independent groups G1 and G2. The left two panels show 95% confidence intervals, with overlap 0.5 and 0, and the corresponding approximate P-values, 0.05 and 0.01 respectively. (Overlap of 0.5 means overlap equal to half the length of one arm of either confidence interval.) The right two panels show standard error bars, with gap 1 and 2 standard errors, and corresponding approximate P-values, 0.05 and 0.01 respectively (one standard error is one arm of the standard error bar). The vertical scale at right shows the P-value as a function of the G2 mean, all other aspects remaining the same. This scale, which applies to all four panels, shows that the P-values stated are a little conservative - the precise values in this case are a little lower than 0.05 and 0.01

Figure 8. Hypothetical data representing a design that includes a repeated measure. The error bars (whether they are confidence intervals or standard errors) can only be used to draw inferences about control and treatment means at the same measurement time and cannot be used to infer temporal differences or trends. The guidelines of Figure 7 may be used to make inferences about independent groups, for example Control vs Treatment means at Time 1, but may not be used to compare a repeated measure, such as Treatment mean at Time 1 vs Treatment mean at Time 2.

Figure 1

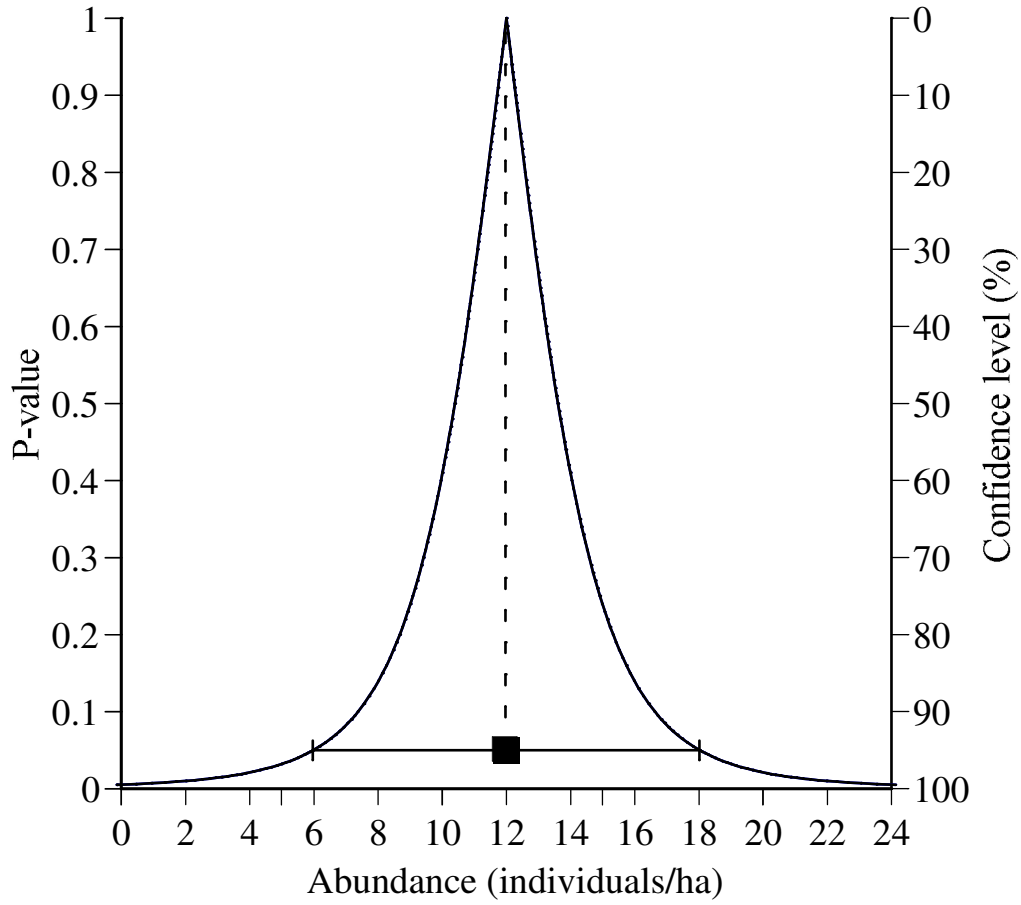


Figure 2

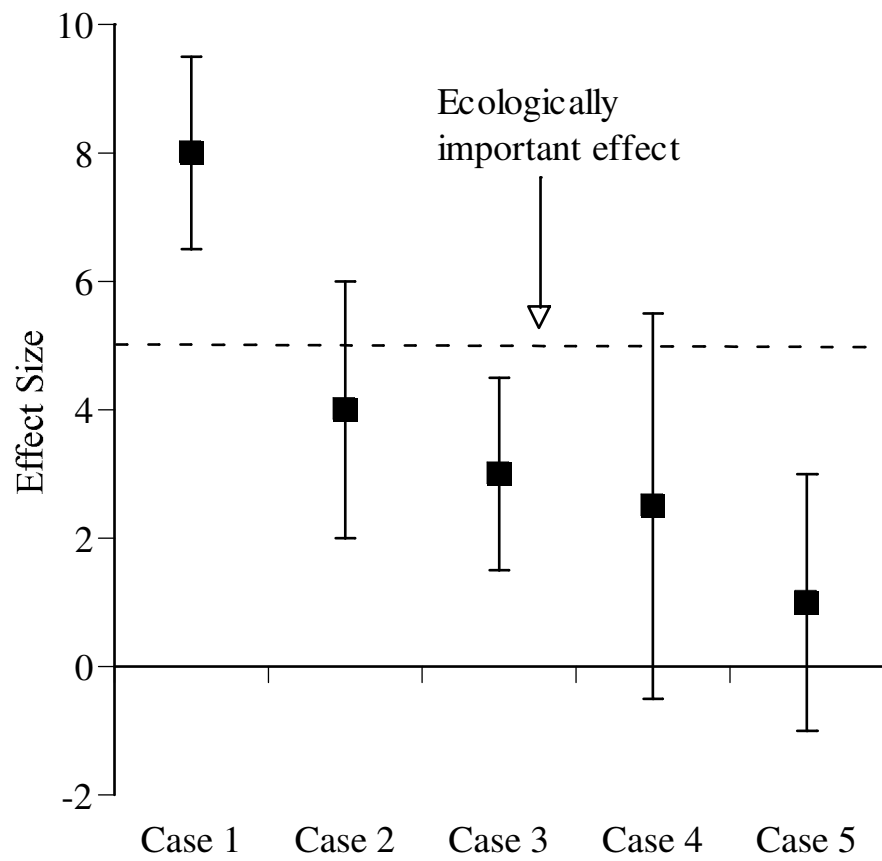


Figure 3

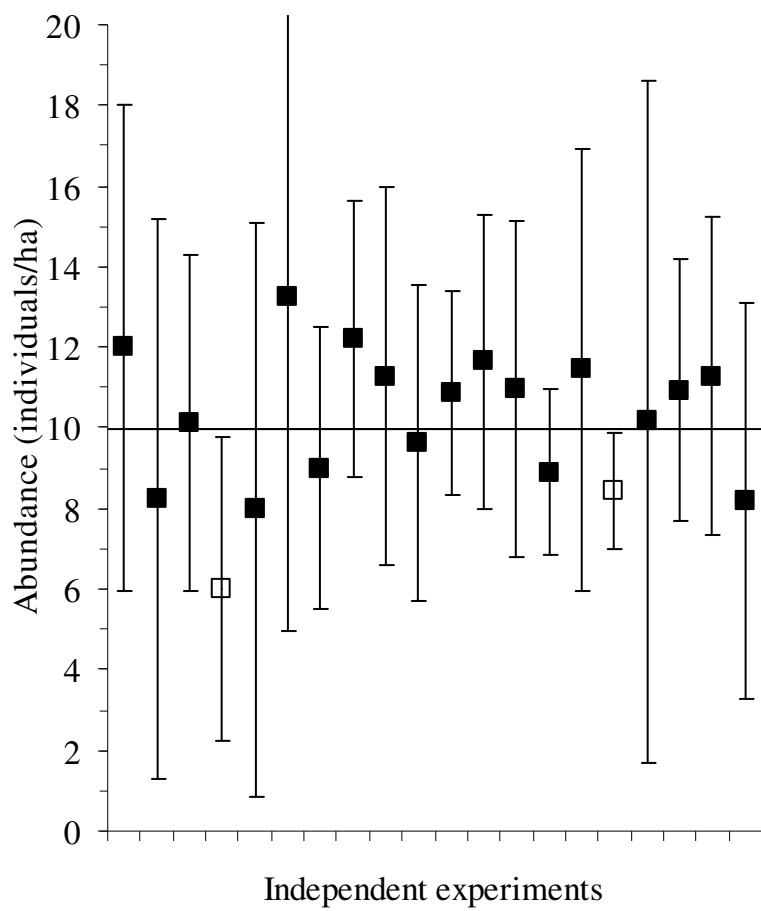


Figure 4

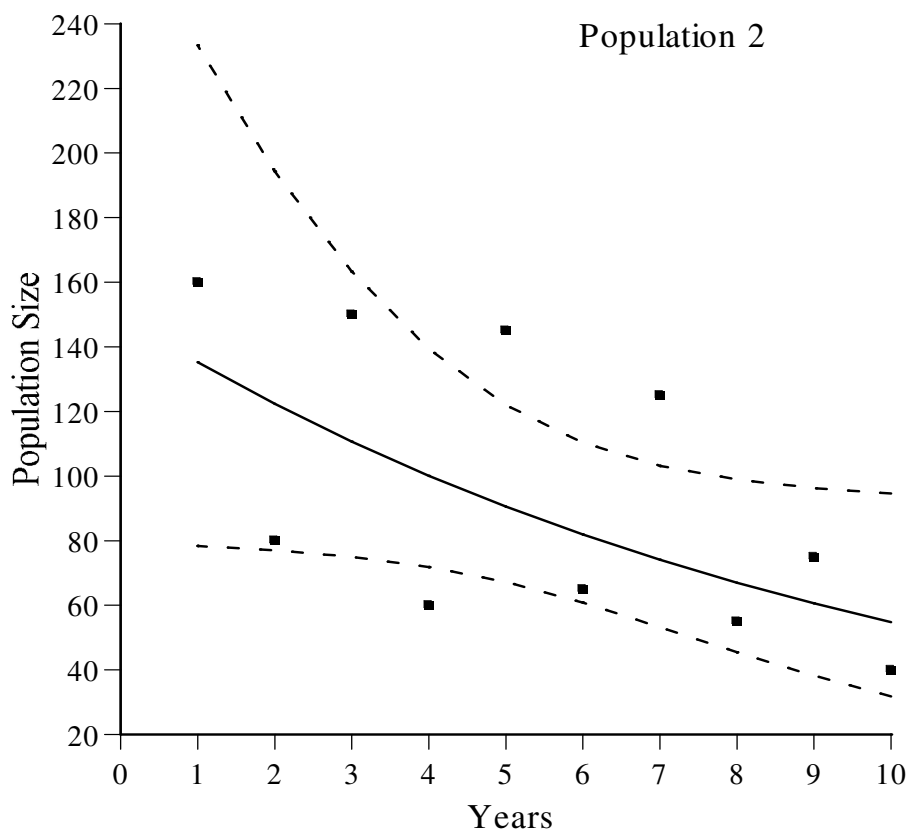
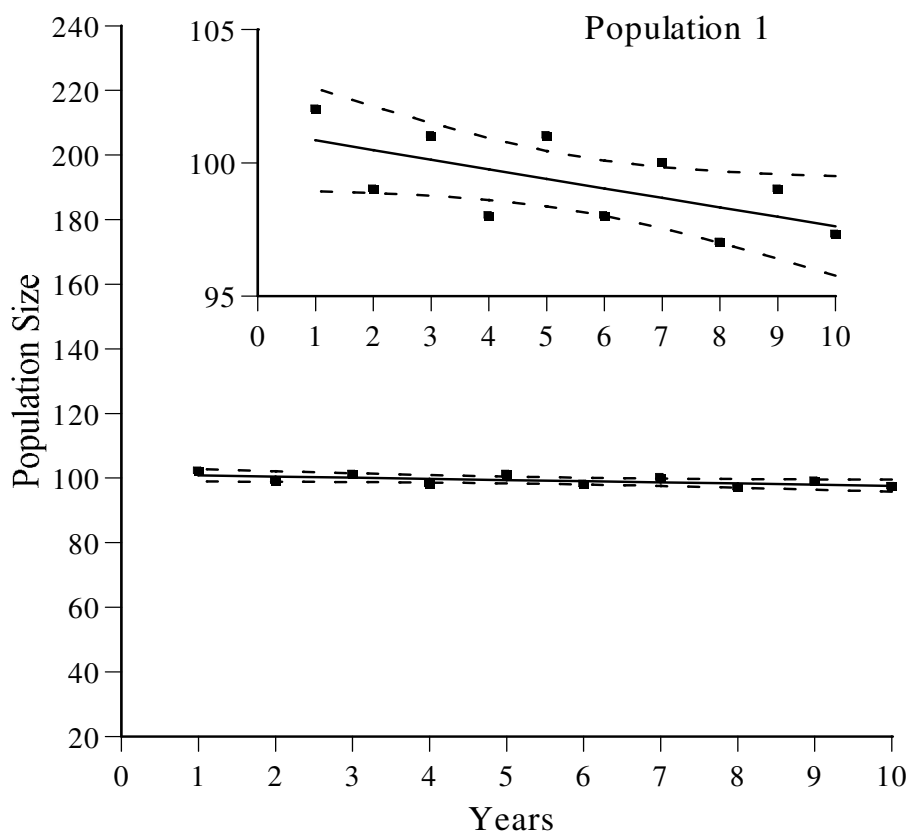


Figure 5

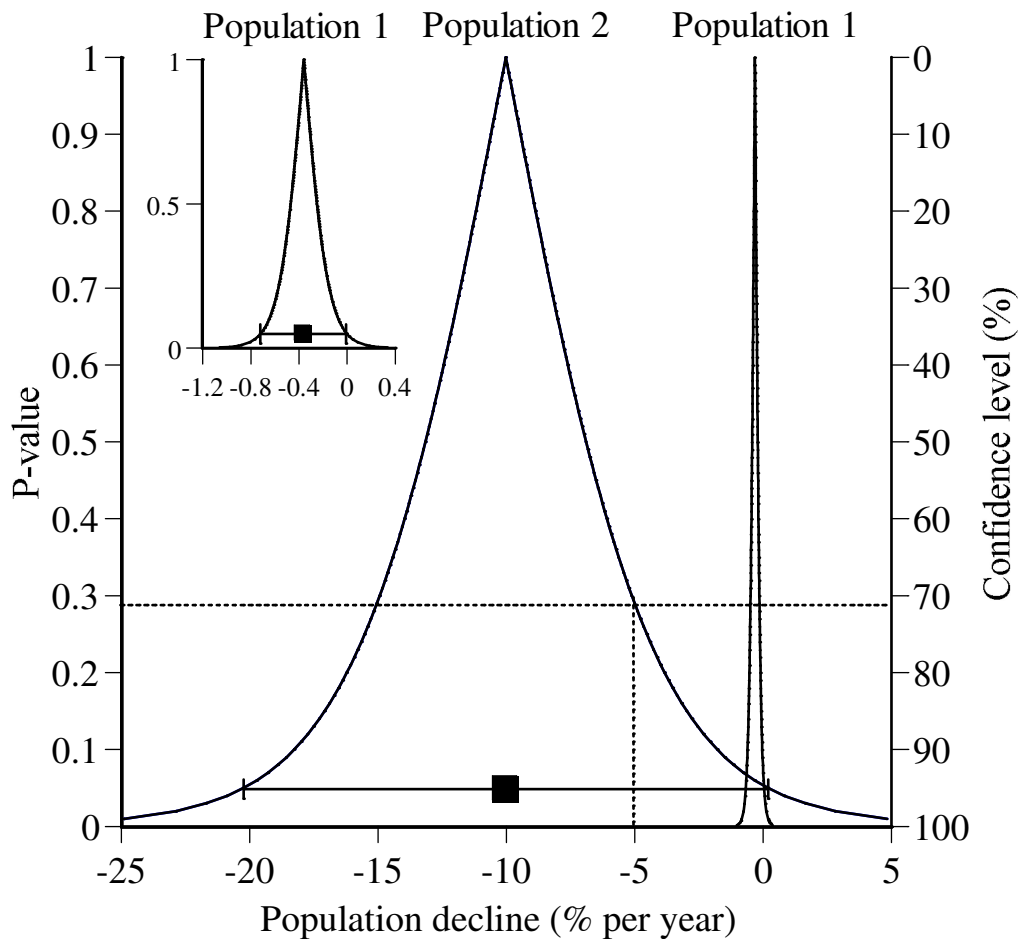


Figure 6

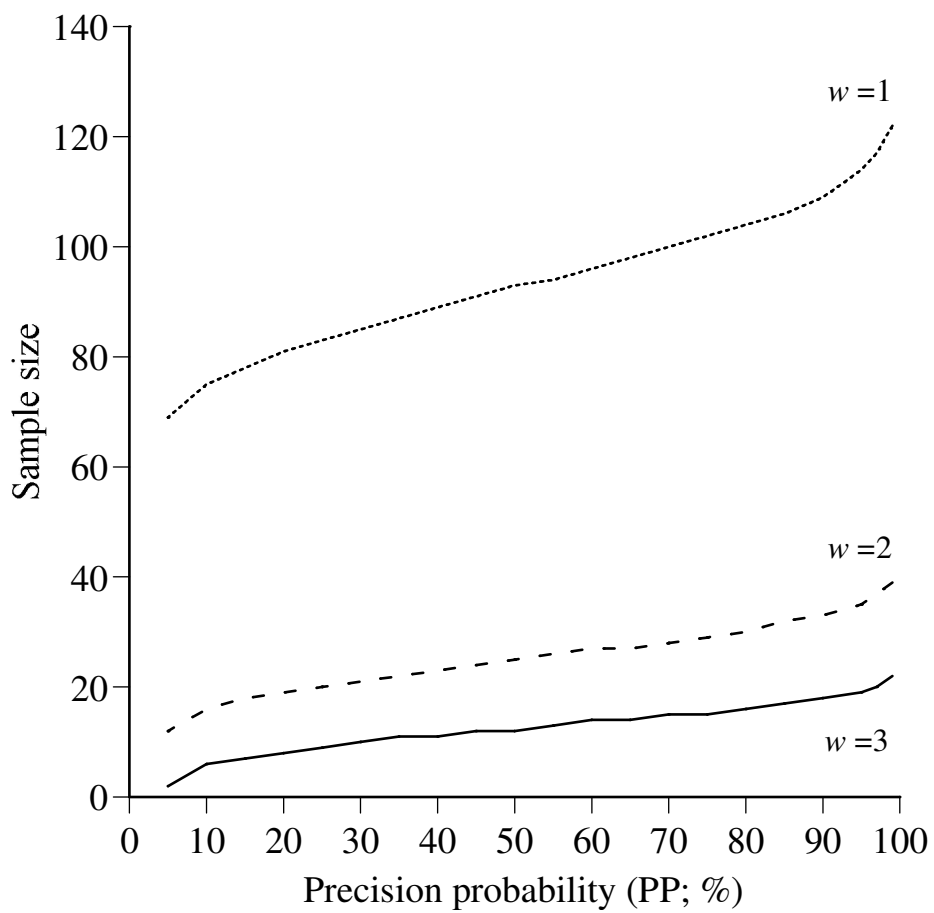


Figure 7.

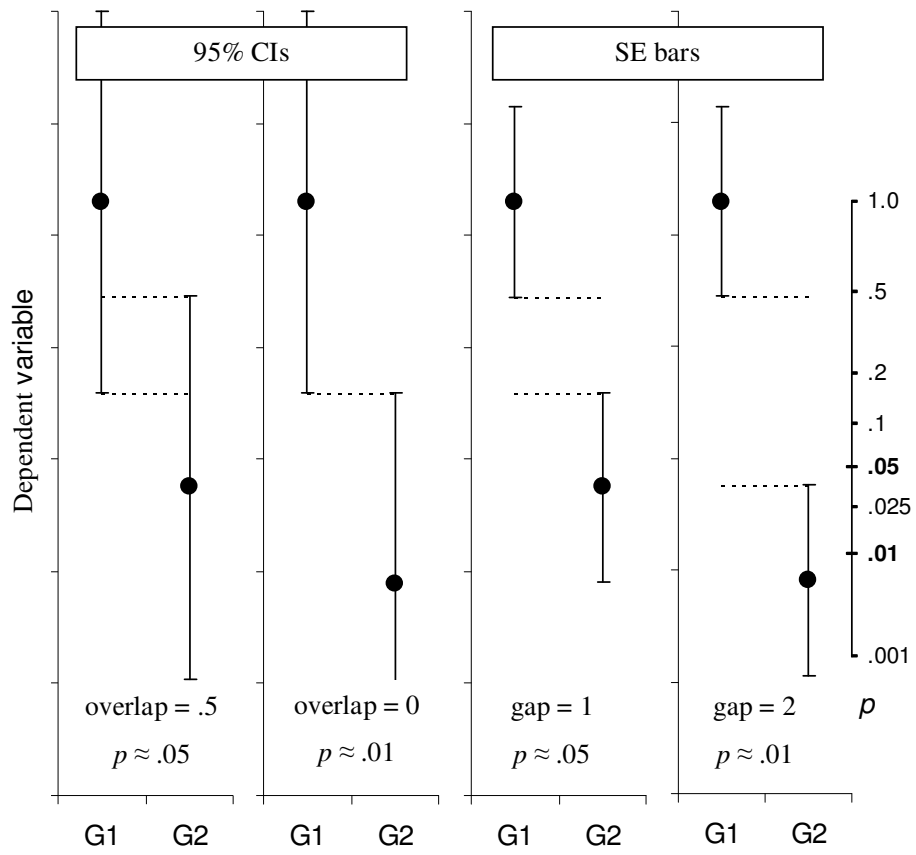
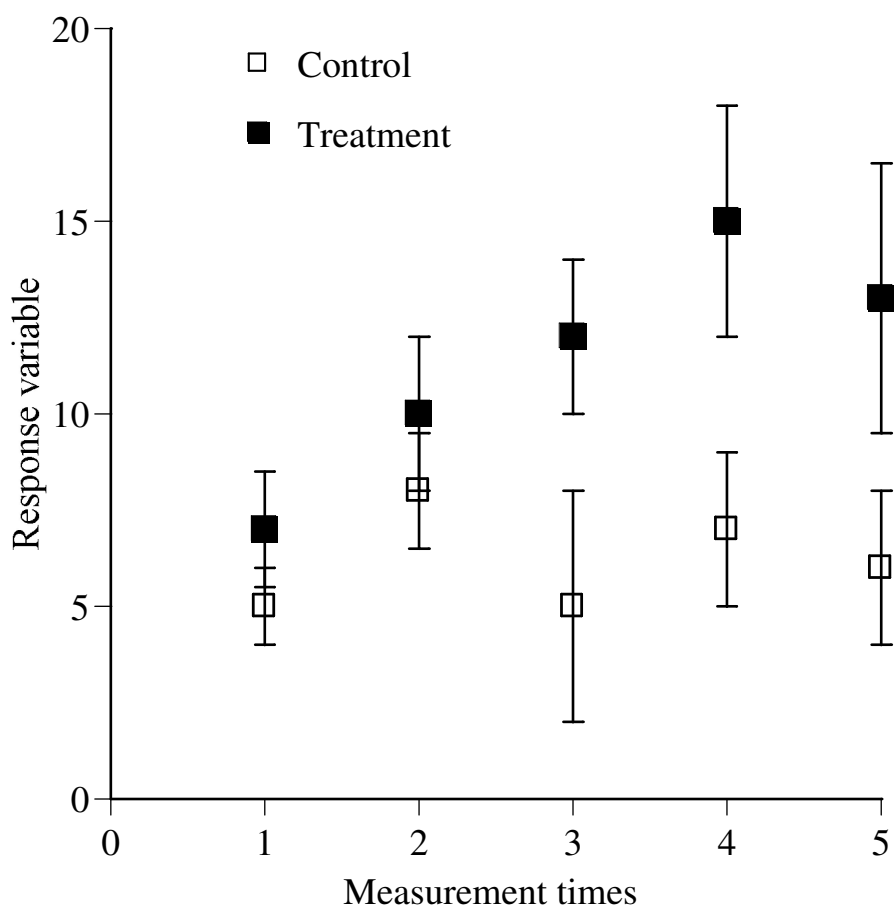


Figure 8



References

- Algina J., Moulder B. C. & Moser B. K. (2002) Sample size requirements for accurate estimation of squared semi-partial correlation coefficients. *Multivariate Behavioral Research* 37: 37-57.
- Algina J. & Olejnik S. (2000) Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research* 35: 119-136.
- Anderson D. R., Burnham K. P. & Thompson W. L. (2000) Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* 64: 912-923.
- Anderson D. R., Link W. A., Johnson D. H. & Burnham K. P. (2001) Suggestions for presenting the results of data analyses. *Journal of Wildlife Management* 65: 373-378.
- Barrick K. A. (2003) Comparison of the nutrient ecology of coastal *Banksia grandis* elfinwood (windswept shrub-like form) and low trees, Cape Leeuwin-Naturaliste National Park, Western Australia. *Austral Ecology* 28: 252-262.
- Bayes T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53: 370-418.
- Belia S., Fidler F., Williams J. & Cumming G. (submitted) Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*.

Berger J. O. & Sellke T. (1987) Testing a point null hypothesis - the irreconcilability of P-values and evidence. *Journal of the American Statistical Association* 82: 112-122.

Burnham K. P. & Anderson D. R. (1998) *Model Selection and Inference: a Practical Information-Theoretic Approach*. Springer, New York.

Burnham K. P. & Anderson D. R. (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* 28: 111-119.

Chatfield C. (1985) The initial examination of data. *Journal of the Royal Statistical Society, Series A* 148: 214-253.

Cohen J. (1994) The earth is round ($p < .05$). *American Psychologist* 49: 997-1003.

Cox D. R. (1977) The role of significance tests. *Scandinavian Journal of Statistics* 4: 49-70.

Cumming G. & Finch S. (submitted) Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist*.

Cumming G., Williams J. & Fidler F. (2004) Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics (in press)*.

- Daly L. E. (2000) Confidence intervals and sample sizes. In: *Statistics with Confidence* (eds. D. G. Altman, D. Machin, T. Bryant, N. & M. J. Gardner) pp. 139-152. BMJ Books, Bristol.
- Di Stefano J. (2001) Power analysis and sustainable forest management. *Forest Ecology and Management* 154: 141-153.
- Di Stefano J. (2003) How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology* 17: 707-709.
- Di Stefano J. (2004) A confidence interval approach to data analysis. *Forest Ecology and Management* 187: 173-183.
- Downes B. J., Barmuta L. A., Fairweather P. G., Faith D. P., Keough M. J., Lake P. S., Mapstone B. D. & Quinn G. P. (2002) *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*. Cambridge University Press, Cambridge.
- Ellison A. M. (1996) An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6: 1036-1046.
- Ellison A. M. (2004) Bayesian inference in ecology. *Ecology Letters* 7: 509-520.
- Fairweather P. G. (1991) Statistical power and design requirements for environmental monitoring. *Australian Journal of Marine and Freshwater Research* 42: 555-567.

Fidler F., Cumming G., Burgman M. & Thomason N. (in press) Statistical reform in medicine, psychology and ecology. *Journal of Socio-economics*.

Fidler F., Thomason N., Cumming G., Finch S. & Leeman J. (2004) Editors can lead researchers to confidence intervals, but can't make them think - Statistical reform lessons from medicine. *Psychological Science* 15: 119-126.

Fisher R. A. (1925) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

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

Fisher R. A. (1955) Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B* 17: 69-77.

Fisher R. A. (1956) *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

Foster J. R. (2001) Statistical power in forest monitoring. *Forest Ecology and Management* 151: 211-222.

Fox D. R. (2001) Environmental power analysis - a new perspective. *Environmetrics* 12: 437-449.

Gerard P. D., Smith D. R. & Weerakkody G. (1998) Limits of retrospective power analysis. *Journal of Wildlife Management* 62: 801-807.

Gerrodette T. (1993) Trends: software for a power analysis of linear regression. *Wildlife Society Bulletin* 21: 515-516.

Gigerenzer G. (1993) The superego, the ego, and the id in statistical reasoning. In: *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (eds. G. Keren & C. Lewis) pp. 311-339. Lawrence Erlbaum, Hillsdale.

Gigerenzer G., Swijtink Z., Porter T., Daston L., Beatty J. & Kruger L. (1989) *The empire of chance. How probability changed science and everyday life*. Cambridge University Press, Cambridge.

Goodman S. N. & Berlin J. A. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine* 121: 200-206.

Hoening J. M. & Heisey D. M. (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 55: 19-24.

Hubbard R. & Ryan P. A. (2000) The historical growth of statistical significance testing in psychology - and its future prospects. *Educational and Psychological Measurement* 60: 661-681.

Johnson D. H. (1999) The insignificance of statistical significance testing. *Journal of Wildlife Management* 63: 763-772.

Johnson D. H. (2002) The role of hypothesis testing in wildlife science. *Journal of Wildlife Management* 66: 272-276.

Keough M. J. & Mapstone B. D. (1997) Designing environmental monitoring for pulp mills in Australia. *Water Science and Technology* 35: 397-404.

Lenth R. V. (2000) Java applets for power and sample size.

<http://www.stat.uiowa.edu/~rlenth/Power/index.html> (but regression module currently found at <http://www.stat.uiowa.edu/~rlenth/Power/OldPiFace.html>).

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

Meehl P. E. (1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46: 806-834.

Nelder J. A. (1999) From statistics to statistical science. *Journal of the Royal Statistical Society Series D - The Statistician* 48: 257-267.

Newcombe R. G. & Altman D. G. (2000) Proportions and their differences. In: *Statistics with Confidence* (eds. D. G. Altman, D. Machin, T. Bryant, N. & M. J. Gardner) pp. 45-56. BMJ Books, Bristol.

Neyman J. (1950) *First Course in Probability and Statistics*. Holt, New York.

Neyman J. (1955) The problem of inductive inference. *Communications on Pure and Applied Mathematics* 8: 13-46.

Neyman J. (1957) Inductive behaviour as a basic concept of philosophy of science. *International Statistical Review* 25: 7-22.

Neyman J. & Pearson E. S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* 20A: 175-240.

Oakes M. W. (1986) *Statistical inference: A commentary for the social and behavioural sciences*. John Wiley and Sons, Chichester.

Osenberg C. W., St Mary C. M., Schmitt R. J., Holbrook S. J., Chesson P. & Byrne B. (2002) Rethinking ecological inference: density dependence in reef fishes. *Ecology Letters* 5: 715-721.

Palmer A. R. (2000) Quasireplication and the contract of error: Lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics* 31: 441-480.

Pearson E. S. (1962) Some thoughts on statistical inference. *Annals of Mathematical Statistics* 33: 394-403.

Peterman R. M. (1990) Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Science* 47: 2-15.

Quinn G. P. & Keough M. J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Rosnow R. L. & Rosenthal R. (1989) Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* 44: 1276-1284.

Rothman K. J. (2002) *Epidemiology: An Introduction*. Oxford University Press, New York.

Saville D. J. (2003) Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology* 57: 167-175.

Schenker N. & Gentleman J. F. (2001) On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician* 55: 182-186.

Schmidt F. L. (1996) Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1: 115-129.

Schmidt F. L. & Hunter J. E. (1997) Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: *What if there were no significance tests?* (ed. J. H. Steiger) pp. 37-64. Lawrence Erlbaum, Mahwah.

Sellke T., Bayarri M. J. & Berger J. O. (2001) Calibration of P-values for testing precise null hypotheses. *American Statistician* 55: 62-71.

Senn S. (1997) *Statistical issues in drug development*. John Wiley & Sons, Chichester.

Smithson M. (2003) *Confidence Intervals*. Sage, Thousand Oaks.

Sokal R. R. & Rohlf F. J. (1995) *Biometry*. W.H. Freeman and Company, New York.

Steidl R. J., Hayes J. P. & Schaubert E. (1997) Statistical power analysis in wildlife research. *Journal of Wildlife Management* 61: 270-279.

Steidl R. J. & Thomas L. (2001) Power analysis and experimental design. In: *Design and Analysis of Ecological Experiments* (eds. S. M. Scheiner & J. Gurevitch) pp. 14-36. Oxford University Press, New York.

Stewart-Oaten A. (1995) Rules and judgments in statistics: Three examples. *Ecology* 76: 2001-2009.

Taylor B. L. & Gerrodette T. (1993) The uses of statistical power in conservation biology: The Vaquita and Northern Spotted Owl. *Conservation Biology* 7: 489-500.

Thomas L. & Juanes F. (1996) The importance of statistical power analysis: an example from *Animal Behaviour*. *Animal Behaviour* 52: 856-859.

Tukey J. W. (1977) *Exploratory data analysis*. Addison-Wesley, Reading.

Tversky A. & Kahneman D. (1971) Belief in the law of small numbers. *Psychological Bulletin* 76: 105-110.

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

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

Wade P. R. (2000) Bayesian methods in conservation biology. *Conservation Biology* 14: 1308-1316.

Wilkinson L. (1999) Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* 54: 594-604.

Wolfe R. & Hanley J. (2002) If we're so different why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal* 166: 65-66.

Yoccoz N. G. (1991) Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* 72: 106-111.

Zar J. H. (1999) *Biostatistical Analysis*. Prentice Hall, Upper Saddle River.