Available online at www.sciencedirect.com
SCIENCE DIRECT®
Forest Ecology and Management 187 (2004) 173–183
Forest Ecology and Management
www.elsevier.com/locate/foreco

A confidence interval approach to data analysis

Julian Di Stefano*
Forest Science Centre, Creswick 3363, Australia

Received 11 February 2003; received in revised form 17 March 2003; accepted 1 July 2003

Abstract

The objective of ecological experiments is often to determine whether observed effects are large enough to be ecologically important. Despite this, effect size measures and associated measures of precision are frequently missing in published ecological research. In many cases, *P*-values are the only information available with which to assess the ecological importance of observed effects, but they provide a poor means of assessment. It is argued that specifying an important effect size a priori and then presenting observed effects with their associated confidence intervals is often a more informative way of presenting ecological data. A hypothetical data set is analysed and interpreted using both *P*-values and confidence intervals and the results from these two approaches compared. Effects interpreted using *P*-values were either statistically significant or not, while confidence intervals provided information about statistical significance, the precision of the estimates, and produced a range within which values for the true effects might plausibly lie. The results show that both statistically significant (<0.001) and non-significant (0.100) *P*-values did not provide useful information about the importance of their associated effects. The capacity to use confidence intervals for analysing complex ANOVA designs is discussed and the implications of different data analysis and presentation techniques for forest management are considered.

Keywords: Confidence intervals; *P*-values; Treatment effect; Effect size; Power analysis

1. Introduction

Sokal and Rohlf (1995) define biometry as 'the application of statistical methods to the solution of biological problems'. This is an appropriate definition as it implies that, as ecologists, our primary focus is to investigate patterns and processes occurring in nature. The mechanics of statistical procedures is not (at least for most of us) of primary interest; we simply use statistics as a tool to help derive ecological meaning from our data. This being the case, it is critical that statistical outputs can be interpreted in the context of the ecological questions and hypotheses under investigation. For example, statistical results should include measures of effect and their uncertainties (Nelder, 1999) so that the ecological importance of effects can be determined. Nevertheless, it is common within ecology journals for experimental results to be represented by *P*-value alone without effect size estimates or associated measures of precision (Anderson et al., 2000; Fidler et al., in preparation). Although *P*-values can be used to define statistical significance, they convey little information about the ecological importance of observed treatment effects.

The majority of published ecological experiments are conducted within the frequentist hypothesis testing framework. Within this framework, observations of the natural world lead to the development of theories,

* Tel.: +61-3-5321-4259; fax: +61-3-5321-4277.
E-mail address: julian.distefano@dse.vic.gov.au (J. Di Stefano).

0378-1127/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1127(03)00331-1

which in turn lead to the construction of research and null hypotheses (see Underwood, 1990 for a detailed description of this approach). In the vast majority of cases, research hypotheses propose the occurrence of an effect (or change or impact) while null hypotheses specify that the effect does not exist, or, more precisely, equals zero. Data are collected and exposed to formal statistical tests that, amongst other outputs, produce statements (*P*-values) concerning the probability of the collected data (or data more extreme) assuming that the null hypothesis is true. Correctly interpreted, small *P*-values (conventionally ≤ 0.05) indicate that observed treatment effects are (probably) larger than zero while large *P*-values (>0.05) suggest there is insufficient statistical evidence to reject the null hypothesis.

Although *P*-values provide a means for determining whether an effect exists they say little about the magnitude of an effect (Jaccard and Guilamo-Ramos, 2002). What ecologists (and scientists in other fields) generally want to know is whether a treatment effect is large enough to be important in the context of the ecological system under investigation. Thus for ecologists, it is important to differentiate between statistically significant and ecologically important treatment effects (Yoccoz, 1991; Steidl et al., 1997; Fox, 2001), an issue that is rarely discussed in published ecological research (Anderson et al., 2000; Fidler et al., in preparation). In many situations, differentiating between statistical significance and ecological importance can be achieved by calculating confidence intervals around observed treatment effects.

The confidence interval approach to data analysis advocated in this paper focuses on estimation of treatment effects and their associated errors (Steidl et al., 1997; Johnson, 1999) and on the specification of an important effect size before the experiment begins. To better understand the meaning of confidence intervals, imagine that an experiment was conducted a large number of times and a 95% confidence interval for the treatment effect was constructed on each occasion. On average, 95% of these intervals will contain the true (population) treatment effect (Levine and Ensom, 2001). A single 95% confidence interval does not contain the true treatment effect with 95% certainty, but can be interpreted as representing a range within which the true effect may plausibly lie (Hoenig and Heisey, 2001; Steidl and Thomas, 2001). Confidence intervals provide an estimate of the true size of treatment effects, and thus can be used to assess the ecological importance of an observed effect based on a single sample of data. In many situations, confidence intervals are completely compatible with traditional hypothesis testing procedure—if a 95% confidence interval does not include the value specified by the null hypothesis (usually zero), the null hypothesis can be rejected at the 5% level (Steidl and Thomas, 2001).

If researchers can define an ecologically important treatment effect a priori (an issue which is discussed in more detail later), the information provided by confidence intervals can lead to one of the five alternative conclusions that differentiate between statistical significance and ecological importance (Fig. 1). While the use of confidence intervals has been advocated in a number of research fields (e.g. wildlife biology, medical science and psychology), discussion of this topic and use of the techniques is all but absent in many mainstream ecological journals. A survey of 45 papers recently published in *Forest Ecology and Management* (172 (2–3) to 173 (1–3) inclusive) found that although the detection of differences or trends was important in 42 papers, confidence intervals around effects was only presented in five.

The objective of this paper is to introduce the concepts associated with a confidence interval approach to data analysis and to generate debate regarding its utility. The discussion is conducted within a frequentist statistical framework; other data analysis approaches (e.g. Bayesian or information theoretic, see Ellison, 1996; Anderson et al., 2000 for an outline of these techniques) are not considered. An example based on a frequently used experimental design (two-way fixed factor ANOVA) was employed to evaluate results based on both *P*-values and confidence intervals. The capacity to use confidence intervals to analyse more complex ANOVA designs is discussed and the implications of different data analysis and presentation methods for forest management are considered.

2. Methods

The consumption of plantation seedlings by mammalian herbivores is problematic in many parts of the world and a number of studies (e.g. Montague, 1993;

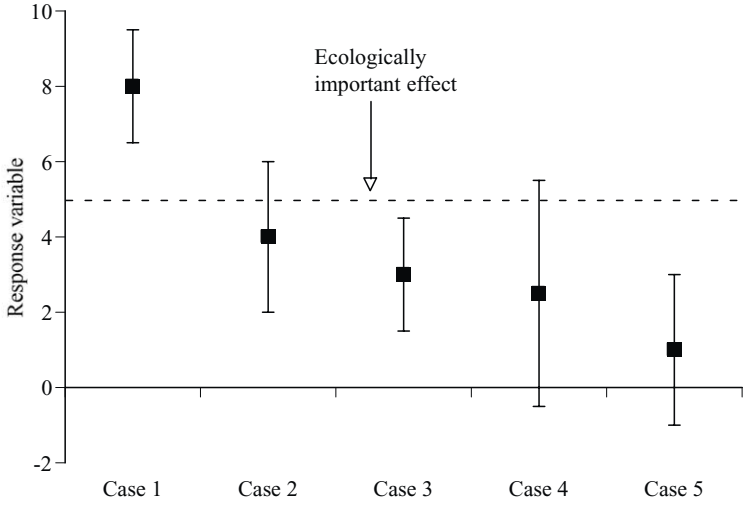

Fig. 1. Interpretation of results using confidence intervals. Black squares are observed treatment effects, error bars are 95% confidence intervals around these effects and the dashed line represents an (arbitrarily defined) ecologically important effect. In Case 1, the observed effect is both statistically 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 and Thomas (2001).

Stange and Shea, 1998; Delisle, 1999; Dubois et al., 2000) have tested the effectiveness of tree guards as a browsing reduction strategy. Based on these studies, a hypothetical data set was generated and analysed, and results interpreted using both *P*-values and 95% confidence intervals. Although the data were hypothetical and structured to support particular arguments, data of similar form are commonly observed in the ecological literature.

The objectives of this hypothetical experiment were to determine: (a) if browsing by mammalian herbivores reduced the height of 1-year-old plantation grown eucalypt seedlings to an important degree, (b) if tree guards provided adequate protection against browsing damage and (c) whether browsing damage or the degree of protection afforded by the tree guards varied in space. Three treatments (unprotected, short tree guards and tall tree guards) and a control (fenced seedlings) were established when seedlings were planted and replicated three times at each of two sites. Seedling height (the response variable) was recorded after 1 year of growth. The data were analysed using a two-way analysis of variance (ANOVA), where Factor A (treatment) had four levels, Factor B (site) had two levels and both factors are considered fixed. In this example Factor B was considered fixed because in many large-scale forestry experiments operational constraints prevent sites from being chosen at random. ANOVA was used as an example because this analysis technique is common in the ecological literature—24 out of 45 papers published in volumes 172 (2–3) to 173 (1–3) of *Forest Ecology and Management* used ANOVA or *t*-tests for data analysis. The analysis was performed using Genstat 5 and the probability of making a Type I error (α) was set at 0.05. Assumptions of normality and homogeneity of variance were tested using normal plots of standardised residuals and plots of standardised residuals versus fitted values, respectively, and no transformations were deemed necessary. Subsequent to the initial ANOVA procedure, *P*-values and 95% confidence intervals were calculated for a number of selected contrasts using Genstat's COMPARISON function.

Based on discussions with experienced researchers (J. Bulinski and C. McArthur, personal communication), a reduction in seedling height of $\geq$30% relative to control seedlings was chosen, a priori, as an important treatment effect, as this degree of height loss after one year was considered likely to have adverse commercial implications. This effect size was also

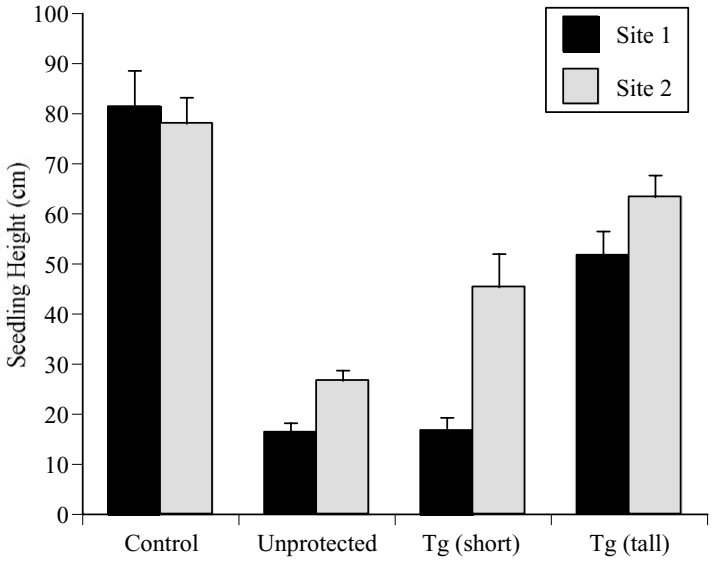

Fig. 2. Treatment means (+S.E.) for all treatments at each site Tg = tree guard.

used to specify an important interaction effect as it would represent substantial differences in the effect of treatments at each site. Because the height of control seedlings was similar at both sites (81.3 and 78.0 cm, see Fig. 2), the important degree of height loss (≥23.9 cm) was calculated using pooled data. Although other effect size measures can be calculated (e.g. Fidler and Thompson, 2001) raw mean differences were used as the effect size in this paper because they are easier to interpret than other indices.

3. Results and discussion

The data are presented graphically in Fig. 2 and results of the original ANOVA are displayed in Table 1. Selected contrasts exploring the interaction term and treatment effects at sites 1 and 2 are shown in Table 2.

Table 1
ANOVA table describing the original analysis

Source	d.f.	SS	MS	*F*	*P*-value
Site	1	937.5	937.5	13.3	0.002
Treatment	3	12666.8	4222.3	59.7	<0.001
Site × treatment	3	780.2	260.1	3.7	0.035
Residual	16	1132.0	70.8		
Total	23	15516.5			

The first three lines of Table 2 represent interaction contrasts that compare the difference between the heights of the control and treatment seedlings at site 1 with the corresponding differences at site 2. For example, the control–unprotected × site contrast compares the difference between the mean height of control and unprotected seedlings at site 1 ($81.3 - 16.3 = 65$ cm) with the corresponding difference at site 2 ($78.0 - 26.7 = 51.3$ cm). The difference between the two ($65 - 51.3 = 13.7$ cm) is the effect listed in Table 2. The non-significant *P*-value associated with this effect indicates that the null hypothesis (browsing reduces seedling height by the same amount at both sites) cannot be rejected. The remainder of Table 2 lists contrasts exploring control–treatment differences at each site. The corresponding effect sizes represent seedling height loss for each treatment relative to the controls. Smaller numbers indicate more effective treatments. The results are described and discussed in terms of conventional statistical interpretation and how they would be interpreted using confidence intervals.

3.1. Conventional statistical interpretation

The initial analysis (Table 1) showed statistically significant treatment, site and interaction effects. The interaction contrasts (Table 2) indicated that the degree

Table 2
Selected contrasts exploring the interaction term from the original ANOVA and treatment effects at sites 1 and 2[a]

Contrast	Effect size (cm)	95% CI	S.E.	*F*	*P*-value
Interaction contrasts					
Control–unprotected × site	13.7	−15.4 to 42.8	13.74	2.0	0.179
Control–short tree guard × site	32.0	2.9 to 61.1	13.74	10.9	0.005
Control–tall tree guard × site	17.6	−11.5 to 46.7	13.74	3.3	0.088
Relevant contrasts at site 1[b]					
Control–unprotected	65.0	50.4 to 79.6	6.87	89.6	<0.001
Control–short tree guard	64.7	50.1 to 79.3	6.87	88.7	<0.001
Control–tall tree guard	29.7	15.1 to 44.3	6.87	18.7	<0.001
Relevant contrasts at site 2[b]					
Control–unprotected	51.3	36.7 to 65.9	6.87	55.9	<0.001
Control–short tree guard	32.7	18.1 to 47.3	6.87	22.6	<0.001
Control–tall tree guard	12.0	−2.6 to 26.6	6.87	3.1	0.100

[a] Effect sizes for the interaction contrasts represent the consistency of the control–treatment differences between sites. Effect sizes for the individual site contrasts represent height loss relative to the controls. The predetermined important effect size is 23.9 cm.

[b] *F*-ratios, standard errors and *P*-values for the individual site contrasts are based on the mean square residual and associated degrees of freedom from the original data set (Quinn and Keough, 2002).

to which short tree guards reduced browsing damage relative to the controls differed between sites. Because of the problems associated with interpreting main effects when the interaction is statistically significant (Underwood, 1997), treatment effects were explored separately at each site using the mean square residual from the original data set (Table 1) to calculate *F*-ratios and *P*-values (Quinn and Keough, 2002). The individual site contrasts (Table 2) indicated that all treatments except for tall tree guards at site 2 resulted in statistically significant reductions in seedling height relative to the controls. In other words, mammalian browsing caused a statistically significant reduction in seedling height at both sites while the only tree guard treatment that prevented a statistically significant reduction in seedling height were tall tree guards at site 2.

3.2. Statistical interpretation using confidence intervals

The 95% confidence intervals around all effects are presented in Table 2. In general the intervals are wide indicating that the effect size estimates are imprecise. For the interaction contrasts all the confidence intervals contain the predetermined important effect (23.9 cm) which suggests that the true (population) effects estimated by these contrasts may or may not be important. Even without this predetermined important effect size the extremes of the intervals indicate that the true interaction effect may be either insubstantial or large. If information about the interaction were of primary importance (as it is in many ecological experiments) more data would be needed so interaction effects could be estimated with greater precision.

The analysis of treatment effects at each site (Table 2) showed that the pre-defined important effect (23.9 cm) is less than the lower 95% confidence bound for the control–unprotected contrasts at both sites. This indicates that browsing reduces the height of seedlings by more than the critical 30% margin. The result is the same for the control–short tree guard contrast at site 1 where the conclusion is that short tree guards are an ineffective browsing reduction strategy. The results for all the other individual site contrasts (control–tall tree guards at both sites and control–short tree guards at site 2) are unclear. In these cases the predefined important effect lies within the confidence intervals thus it is plausible that true height reductions could be either greater or less than the critical 30% figure. Consequently, the results are inconclusive and more data are required before the impact of these treatments can be determined.

A summary of conclusions corresponding to the analysis of individual site contrasts is presented in Table 3. This comparison shows that *P*-values cannot be used to indicate whether treatment effects are

Table 3
Summary of conclusions based on *P*-values and 95% confidence intervals for the analysis of individual site contrasts

Individual site contrasts	*P*-values	95% CI
Site 1		
Control–unprotected	Unprotected (browsed) seedlings experienced statistically significant (>0) height losses	The height of unprotected (browsed) seedlings was reduced to an important degree
Control–short tree guards	Seedlings protected by short tree guards experienced statistically significant (>0) height losses	Short tree guards did not prevent browsing damage: seedling height losses were important
Control–tall tree guards	Seedlings protected by tall tree guards experienced statistically significant (>0) height losses	Uncertain if tall tree guards prevented browsing damage: seedling height losses may or may not have been important
Site 2		
Control–unprotected	Unprotected (browsed) seedlings experienced statistically significant (>0) height losses	The height of unprotected (browsed) seedlings was reduced to an important degree
Control–short tree guards	Seedlings protected by short tree guards experienced statistically significant (>0) height losses	Uncertain if short tree guards prevented browsing damage: seedling height losses may or may not have been important
Control–tall tree guards	Height losses for seedlings protected by tall tree guards were not statistically significant. Insufficient evidence to reject the null hypothesis (H_0: height difference = 0)	Uncertain if tall tree guards prevented browsing damage: seedling height losses may or may not have been important

important. Although the *P*-values associated with the control–tall tree guard contrast at site 1 and the control–short tree guard contrast at site 2 are small (both <0.001), confidence interval analysis showed that these treatments may or may not have reduced browsing damage by an important degree. This corresponds to Case 2 in Fig. 1 where the effect is statistically significant, but may or may not be important in the context of the experimental objectives. The importance of the treatment effect associated with the control–tall tree guard contrast at site 2 is similarly unclear, even though the *P*-value associated with this effect is relatively large (0.100). This corresponds to Case 4 in Fig. 1 where the effect is not statistically significant but may or may not be important. In all three cases, the data were inadequate to estimate the treatment effects with sufficient precision to determine whether they were important or not.

The inability to detect effects with sufficient precision results from low power statistical tests, an occurrence that is common in the scientific literature (Sedlmeier and Gigerenzer, 1989; Peterman, 1990; Fairweather, 1991). In the case of the present example the conclusion that the statistical tests lacked power is not surprising as each treatment was only replicated three times at each site. Retrospective power analysis can also be used to assess statistical power after an experiment has been conducted (Fairweather, 1991; Quinn and Keough, 2002) but recent criticism of this procedure (Goodman and Berlin, 1994; Steidl et al., 1997; Gerard et al., 1998; Hoenig and Heisey, 2001; Lenth, 2001) has led some authors to recommend using confidence intervals in lieu of retrospective power analysis (Levine and Ensom, 2001; Steidl and Thomas, 2001; Hoenig and Heisey, 2001) although conclusions stemming from both techniques may often be similar (Thomas, 1997). In many situations, problems of low power in ecological field experiments can be avoided (or at least revealed) by conducting a priori power analysis. The utility of this procedure has been discussed in a number of recent publications (Fairweather, 1991; Keough and Mapstone, 1997; Di Stefano, 2001; Foster, 2001; Downes et al., 2002; Quinn and Keough, 2002), and it is clear that incorporating power analysis into the planning stage of ecological experiments will facilitate the generation of more useful statistical results.

3.3. Specification of important treatment effects

Interpretation of results using confidence intervals is facilitated by the a priori specification of important

treatment and interaction effects. Determination of an important effect is often difficult (particularly for interactions) and may, at times, be arbitrary and subjective (Rotenberry and Wiens, 1985). Nevertheless, attempting to specify important effect sizes is critical as it is often meaningless to view effects as important simply because they are larger than zero. The key is to consider the system under investigation and, as best as possible, define an effect size that has important consequences (Keough and Mapstone, 1997; Downes et al., 2002).

One factor that complicates the specification of important effects is the number of different effect size measures—overall more than 60 have been specified in the statistical literature (Jaccard and Guilamo-Ramos, 2002). For example, the effect attributable to a main factor in a two-way ANOVA can be described in terms of a raw effect (e.g. mean difference) or standardised effect sizes such as Cohen's f and proportion of explained variance (PEV) measures including eta squared (η^2) and omega squared (ω^2) (Fidler and Thompson, 2001). When possible, raw effect sizes should be used as they are in the units of the response variable and thus much easier to interpret. A number of additional arguments against the use of standardised effect size measures are outlined by Lenth (2001) and Jaccard and Guilamo-Ramos (2002).

For interactions, however, even raw effect sizes are difficult to interpret. Raw effect sizes for the overall interaction effect can be calculated (Carey, 2002) but the output is generally not helpful. For the example used in this paper, the overall interaction effect is 31.8 cm and is influenced by the degree to which all treatments have differing effects at each site. The approach used in this study was to consider selected aspects of the overall interaction by using interaction contrasts. Even so, the effect sizes are difficult to interpret and this difficulty would increase for more complex examples. A useful alternative to raw effects sizes for specifying overall or other complex interaction effects is one of the PEV indices. Using these indices an interaction effect can be expressed as the proportion of variance it explains. Nevertheless, PEV indices do not always reflect the true magnitude of an effect and some are influenced by arbitrary design decisions (Jaccard and Guilamo-Ramos, 2002). In addition, it is difficult (although possible) to calculate confidence intervals around PEV indices for unbalanced designs (Burdick and Graybill, 1992), a situation that frequently occurs in ecological research. There are no simple solutions to this problem—finding meaningful effect size measures and specifying important effects for interaction terms is a difficult matter.

Although generally not considered in ecological studies, the need to define important effects has occasionally been highlighted (e.g. Keough and Mapstone, 1995, 1997; Downes et al., 2002; Whittier et al., 2002). Whittier et al. (2002) acknowledge that the specification of an important effect is always subject to some degree of judgement, and thus present results in the context of both a smaller and larger effect. Extending this concept, some of the difficulties associated with specifying an important treatment effect might be overcome by specifying an important effect *range*. Although it is often difficult to define an important effect precisely, a range of values (bounded on one side by an effect that is obviously unimportant and on the other by one that is obviously important) may be much more satisfactory. Researchers can then use the limits of this range to determine the true importance of effects, and values within it to specify uncertainty. This method is still subjective, but may be more appealing to researchers who are reluctant to specify the exact magnitude of an important effect.

Although the detection of ecologically important treatment effects is likely to be the objective of pure ecological research, the specification of important effects for applied studies may be influenced by economic, social or legal considerations (Keough and Mapstone, 1995; Mapstone, 1995; Downes et al., 2002). This is exemplified in the present paper where the important treatment effect had an economic basis. In studies with both ecological and human dimensions (for example, an investigation into the quantity and type of trees that should remain after native forest timber harvesting) the specified important treatment effect is likely to be influenced by both ecological and social criteria. In cases like this, the definition of an important treatment effect should be made by researchers in consultation with relevant stakeholders, and the final figure is likely to be a compromise between a number of competing interests. A detailed discussion of the complexities involved with defining an important treatment effect is presented in Downes et al. (2002).

3.4. Confidence interval construction and effect size estimation for complex ANOVA designs: applicability, problems and issues

In most situations confidence intervals can be used to report results for more complex ANOVA designs (Fidler and Thompson, 2001; Bird, 2002; Jaccard and Guilamo-Ramos, 2002). Bird (2002) used confidence intervals around raw and standardised effects of planned contrasts for designs including multi-way fixed factor, random factor, mixed model, nested and repeated measures. Fidler and Thompson (2001) focused on the use of PEV effect size measures and described the use of non-central distributions in confidence interval calculation. Although not the focus of this paper, confidence intervals can be constructed around effect size measures for a range of regression analyses as well.

There are, however, a number of problems associated with confidence interval calculation for ANOVA designs. Although the calculation of confidence intervals around raw effect size measures associated with individual contrasts is straightforward (Jaccard and Guilamo-Ramos, 2002), badly unbalanced designs cause problems (El-Bassiouni and Abdelhafez, 2000) particularly for PEV indices (discussed above). Verrill (1999) also identified problems (as well as a solution) for confidence interval calculation where a blocking factor is used.

Confidence intervals are difficult to interpret when effect size measures are in an unfamiliar scale (Bird, 2002). This often occurs when data are transformed (Stewart-Oaten et al., 1992), but the need for transformations can probably be minimised if randomisation techniques are used. A further criticism (often made in relation to α) is that the level of precision tends to be an arbitrary choice (Robinson and Wainer, 2002). 95% confidence intervals are often used by convention, but other levels of precision may be more appropriate. Yet another issue raised by Wilcox (2002) is that confidence intervals based on parametric statistical techniques can be markedly inaccurate. This criticism, however, is of parametric statistics and not of confidence intervals per se. Wilcox (2002) suggested that a variety of 'modern' methods including the use of trimmed means and bootstrapping techniques can be used to generate more reliable estimates of effect sizes and their precision.

Another serious issue is the difficulty involved in specifying ecologically important effect sizes, a task that involves linking an effect size measure to important change in the real world (Jaccard and Guilamo-Ramos, 2002). Although approximating the size of important effects based on imprecise information is appropriate, guessing an important effect size in the absence of data may result in misleading conclusions. In all branches of ecology there is a need for more research into what constitutes ecologically important change and how this can be represented using biologically interpretable effect size measures.

Finally, most commercially available statistical packages focus on statistical null hypothesis testing and routines that generate confidence intervals are not always readily available. For example, MINITAB 13.31 calculates confidence intervals for fixed factor ANOVA but does not do so for the random factor version of the analysis. Bird (2002) notes that none of the popular statistical packages carry out all of the confidence interval based analysis he discusses. Although modern technology makes the calculation of effect size measures and associated confidence intervals possible, the non-standard nature of these procedures means that some confidence interval based outputs are difficult to produce.

3.5. The value of P

The primary argument in this paper has been that calculating confidence intervals around observed effects provides useful information while *P*-values do not. In many cases reporting effect size measures and confidence intervals render *P*-values unnecessary and in these instances they need not be reported (Robinson and Wainer, 2002). However, there are times when effect size measures and associated confidence intervals provide no useful insight and in these situations the use of *P*-values is appropriate. How, for instance, can the difference between groups in an ordination plot be described in such a way as to convey ecological meaning? What do effect size measures and confidence intervals mean when assessing whether interaction terms should be included or excluded from a complex regression model, or when interpreting a complex interaction in an ANOVA? In such situations *P*-values can be used to assess the statistical significance of effects and acknowledgement made that a meaningful estimation of effect size is difficult.

Correctly interpreted *P*-values are an often uninformative, occasionally useful and largely innocuous number. *P*-values, however, are often misinterpreted and this may have serious implications for scientific conclusions or management decisions based on ecological data. Common mistakes include using the size of the *P*-value to draw conclusions about the importance of an effect (Cohen, 1990; Johnson, 2002) and the belief that *P* is the probability that the null hypothesis is true (Johnson, 1999). These (and other) misinterpretations lead to incorrect or at the very least unfounded conclusions and do not help ecologists derive meaning from their data. Many ecologists (Fidler et al., in preparation; Johnson, 2002) as well as scientists in other fields (e.g. Jaccard and Guilamo-Ramos, 2002) make interpretative mistakes of this kind. When *P*-values are used care should be taken to interpret them correctly.

3.6. Implications for forest management and conservation

Environmental management decisions should be based, whenever possible, on relevant ecological data (Murphy and Noon, 1991). The way that ecologists analyse data and present results, however, can influence data interpretation and use by environmental managers. In forestry, the detection of important effects by experiments and monitoring programs often precipitate management action. In the context of the example used in this paper, the decision to employ tree guards as a browsing reduction strategy should only be made if they could prevent seedling height loss from exceeding an ecologically or economically important margin. To this end, an estimate of effect size, the precision of this estimate and some consideration about what constitutes an important effect is required to make an informed management decision.

The heavy reliance on *P*-values in all fields of science (Nelder, 1999) means that managers are often presented with sub-optimal information. As stated earlier, *P*-values simply provide a statistical assessment of results, and thus do not help managers relate experimental outcomes to the real world. In addition, resource managers may be more likely to misinterpret *P*-values than practising scientists. Peterman (1990) suggests that interpreting large *P*-values as 'no effect' is an error that resource managers (and scientists) often make. If effect size measures and confidence intervals are presented, misinterpretation of results is much more difficult.

Management action or non-action is often based on the results of many studies over a long period of time so untangling the consequences of poor statistical reporting practices may often be impossible. Nevertheless, it is likely that inadequate presentation of research results has lead to sub-optimal and ill-informed forest management decisions, or the absence of decisions when they were warranted. Peterman (1990) provides examples from fisheries ecology demonstrating how the interpretation of statistically non-significant results as 'no effect' led to management inaction and a subsequent decline in fish stocks. This example is exactly analogous to experiments designed to test the sustainability of various forestry practices where statistically non-significant results are taken to mean that the practices examined have no effect on the measured variables and thus should remain unaltered. For example, Simon et al. (2002) examined the effect of clear cutting on the abundance of small mammals and found no statistically significant reductions in abundance on clear-cut relative to burnt sites. Based on these data Simon et al. (2002) concluded that small mammal communities can be maintained at clear-cut sites. Although raw effect sizes can be calculated from the data, no estimates of effect size precision are reported. Consequently, it is impossible to tell how large the treatment effects may plausibly have been given the variance in the data. In addition, there is no discussion about what constituted an important change in mammal abundance, an aspect of the study that had clear implications for forest management. Studies like the one conducted by Simon et al. (2002) are important as they provide forest managers with information about the effect of harvesting on various aspects of forest biodiversity. However, presenting results without clear measures of effect, estimates of precision or discussion regarding ecologically important effect sizes makes it difficult to interpret and use the data in a management context. Presenting results along the lines suggested in this paper would assist the appropriate application of ecological data to management decisions.

4. Conclusion

Relative to *P*-values, the construction of confidence intervals around observed effects often enables a more informed appraisal of ecological data. If an important treatment effect is specified a priori, confidence intervals can be used to differentiate between statistical significance and ecological (economic, social, etc.) importance. Using this technique, researchers are able to estimate a plausible range for observed effects and thus ascertain whether their data provide insight into natural patterns and processes. Presenting effect size estimates and associated confidence intervals, however, is not a data analysis panacea—problems with these techniques do exist and 'naked' *P*-values are still useful in some situations. Nevertheless, there seems little reason to use *P*-values alone for applications where meaningful effect size measures and associated confidence intervals are available.

Acknowledgements

Thanks to Lauren Bennett, Kylie McKenzie and Alan York for reading and commenting on early drafts of this manuscript. Kym Butler and members of the Statistics Reading Group at the Forest Science Centre provided valuable discussion of relevant issues.

References

Anderson, D.R., Burnham, K.P., Thompson, W.L., 2000. Null hypothesis testing: problems, prevalence, and an alternative. J. Wildl. Manage. 64, 912–923.

Bird, K.D., 2002. Confidence intervals for effect sizes in analysis of variance. Educ. Psychol. Measur. 62, 197–226.

Burdick, R.K., Graybill, F.A., 1992. Confidence Intervals on Variance Components. Marcel Dekker, New York.

Carey, J.M., 2002. An investigation of hypothesis testing and power analysis in impact assessment, using case studies of marine infauna. Ph.D. Thesis. Department of Zoology, University of Melbourne.

Cohen, J., 1990. Things I have learned (so far). Am. Psychol. 45, 1304–1312.

Delisle, C., 1999. Treeshelters: a judicious choice for improving red ash growth. For. Chronol. 75, 845–849.

Di Stefano, J., 2001. Power analysis and sustainable forest management. For. Ecol. Manage. 154, 141–153.

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.

Dubois, M.R., Chappelka, A.H., Robbins, E., Somers, G., Baker, K., 2000. Tree shelters and weed control: effect on protection, survival and growth of cherrybark oak seedlings planted on a cutover site. New For. 20, 105–118.

El-Bassiouni, M.Y., Abdelhafez, M.E.M., 2000. Interval estimation of the mean in a two-stage nested model. J. Statist. Comput. Simul. 67, 333–350.

Ellison, A.M., 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecol. Appl. 6, 1036–1046.

Fairweather, P.G., 1991. Statistical power and design requirements for environmental monitoring. Aust. J. Mar. Freshwater Res. 42, 555–567.

Fidler, F., Thompson, B., 2001. Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. Educ. Psychol. Measur. 61, 575–604.

Fidler, F.M., Burgman, M.A., Thomason, N.R., Cumming, G.D., Buttrose, R.W., in preparation. Deficiencies in statistical reporting practices. In Conservation Biology Journals.

Foster, J.R., 2001. Statistical power in forest monitoring. For. Ecol. Manage. 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. J. Wildl. Manage. 62, 801–807.

Goodman, S.N., Berlin, J.A., 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann. Int. Med. 121, 200–206.

Hoenig, J.M., Heisey, D.M., 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am. Statist. 55, 19–24.

Jaccard, J., Guilamo-Ramos, V., 2002. Analysis of variance frameworks in clinical child and adolescent psychology: advanced issues and recommendations. J. Clin. Child Psychol. 31, 278–294.

Johnson, D.H., 1999. The insignificance of statistical significance testing. J. Wildl. Manage. 63, 763–772.

Johnson, D.H., 2002. The role of hypothesis testing in wildlife science. J. Wildl. Manage. 66, 272–276.

Keough, M.J., Mapstone, B.D., 1995. Protocols for designing marine ecological monitoring programs associated with BEK mills. Technical Report No. 11. CSIRO, Canberra.

Keough, M.J., Mapstone, B.D., 1997. Designing environmental monitoring for pulp mills in Australia. Water Sci. Technol. 35, 397–404.

Lenth, R.V., 2001. Some practical guidelines for effective sample size determination. Am. Statist. 55, 187–193.

Levine, M., Ensom, M.H.H., 2001. Post hoc power analysis: an idea whose time has passed? Pharmacotherapy 21, 405–409.

Mapstone, B.D., 1995. Scalable decision rules for environmental impact studies: effect size, type I and type II errors. Ecol. Appl. 5, 401–410.

Montague, T.L., 1993. An assessment of the ability of tree guards to prevent browsing damage using captive swamp wallabies (*Wallabia bicolor*). Aust. For. 56, 145–147.

Murphy, D.D., Noon, B.D., 1991. Coping with uncertainty in wildlife biology. J. Wildl. Manage. 55, 773–782.

Nelder, J.A., 1999. Statistics for the new millennium: from statistics to statistical science. Statistician, Part 2 48, 257–269.

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

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

Robinson, D.H., Wainer, H., 2002. On the past and future of null hypothesis significance testing. J. Wildl. Manage. 66, 263–271.

Rotenberry, J.T., Wiens, J.A., 1985. Statistical power and community wide patterns. Am. Nat. 125, 164–168.

Sedlmeier, P., Gigerenzer, G., 1989. Do studies of statistical power have an effect on the power of studies? Psychol. Bull. 105, 309–316.

Simon, N.P.P., Stratton, C.B., Forbes, G.J., Schwab, F.E., 2002. Similarity of small mammal abundance in post-fire and clearcut forests. For. Ecol. Manage. 165, 163–172.

Sokal, R.R., Rohlf, F.J., 1995. Biometry, 3rd ed. Freeman, New York.

Stange, E.E., Shea, K.L., 1998. Effect of deer browsing, fabric mats, and tree shelters on *Quercus rubra* seedlings. Restor. Ecol. 6, 29–34.

Steidl, R.J., Thomas, L., 2001. Power analysis and experimental design. In: Scheiner, S.M., Gurevitch, J. (Eds.), Design and Analysis of Ecological Experiments, 2nd ed. Oxford University Press, New York, pp. 14–36.

Steidl, R.J., Hayes, J.P., Schauber, E., 1997. Statistical power analysis in wildlife research. J. Wildl. Manage. 61, 270–279.

Stewart-Oaten, A., Bence, J.R., Osenberg, C.W., 1992. Assessing effects of unreplicated perturbations: no simple solutions. Ecology 73, 1396–1404.

Thomas, L., 1997. Retrospective power analysis. Conserv. Biol. 11, 276–280.

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

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

Verrill, S., 1999. When good confidence intervals go bad: predictor sort experiments and ANOVA. Am. Statist. 53, 38–42.

Whittier, T.R., Paulsen, S.G., Larsen, D.P., Peterson, S.A., Herlihy, A.T., Kaufmann, P.R., 2002. Indicators of ecological stress and their extent in the population of northeastern lakes: a regional-scale assessment. Bioscience 52, 235–247.

Wilcox, R.R., 2002. Understanding the practical advantages of modern ANOVA methods. J. Clin. Child Psychol. 31, 399–412.

Yoccoz, N.G., 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bull. Ecol. Soc. Am. 72, 106–111.