ELSEVIER

Forest Ecology and Management 154 (2001) 141–153

Forest Ecology and Management

www.elsevier.com/locate/foreco

Power analysis and sustainable forest management

Julian Di Stefano*

Forest Science Centre, Water St., Creswick 3363, Australia

Received 7 June 2000; accepted 4 October 2000

Abstract

This paper discusses power analysis in the context of sustainable forest management. It is suggested that a priori power analysis should be formally incorporated into the planning stage of all experiments designed to test whether forestry practices are sustainable. A priori power analysis enables researchers to estimate the probability of making a Type II error (i.e., finding no significant difference when one in fact exists). This information is critical in the statistical assessment of sustainable forestry, as unwittingly accepting a Type II error could result in poor management decisions. In addition, it is proposed that statistical assessments of sustainable forestry objectives can be more relevant if alpha (α) is liberated from its traditional value of 0.05. It is argued that in the context of sustainable forestry, making a Type II error can be more costly than making a Type I error. Consequently, it often makes sense for beta (β) to be small (say 0.05) and α to take on a larger value. In other situations the cost of making a Type I error may be more important, thus a procedure which enables researchers to determine a locally relevant α:β ratio is recommended.

Keywords: Power analysis; Sustainable forest management

1. Introduction

During the last decade, the concept of sustainable forestry has acquired an increasing global focus. A series of international meetings between 1992 and 1995 have precipitated general agreement about how sustainable forestry should be defined, and how it can be measured and assessed. The outcome of these discussions has been to define sustainable forestry in the context of criteria and indicators. The criteria are general principles that express commonly agreed upon objectives of sustainable forestry, while the indicators are designed to assess whether these objectives are being met (Brand, 1997). The most recent set of internationally recognised criteria and indicators regarding sustainable forest practices were presented as part of the Santiago Declaration during 1995.[1]

Although the criteria in the Santiago Declaration clearly outline the objectives of sustainable forestry, the current set of indicators only provides a broad assessment framework without clear standards or baselines (Brand, 1997). While these indicators may be adequate for assessing the objectives of sustainable forestry at regional or national scales, they are too general for use at the scale of the forest management unit (Turner and Lambert, 1997). Consequently, there is likely to be much work in coming years dedicated to the selection of indicators (and the development of

* Tel.: +61-35321-4259; fax: +61-35321-4166.
E-mail address: julian.distefano@nre.vic.gov.au (J. Di Stefano).

[1] A list of international criteria and indicators are presented in Turner and Lambert (1997). Criteria and indicators that have been recognised in Australia are presented in Commonwealth of Australia (1998).

0378-1127/01/$ – see front matter
PII: S0378-1127(00)00627-7

methods and protocols for their measurement) that enable an assessment of sustainable forestry objectives at local scales. A number of recent studies (e.g., McLaren et al., 1998; Burger and Kelting, 1999; Rab, 1999) exemplify that this task is already under way. There is little doubt that at some time in the future, forest managers will have an impressive array of indicators and methods with which the sustainability of local forestry practices can be assessed.

It is also likely that a hypothetico-deductive framework (e.g., Platt, 1964; Popper, 1968; Underwood, 1990) will be used to assess the sustainability of local forestry practices. In other words, data collected about indicators of sustainable forestry will be used to test the null hypothesis that forestry practices have no significant effect on the level of those variables. In effect, assessments of sustainable forestry can be regarded as a form of environmental impact assessment, where the null hypothesis is that some human activity (in this case forestry practices) has no significant detrimental effect on the environment (Fairweather, 1991). Within this framework, statistical procedures will be used to test the null hypothesis of no effect, and the subsequent results will provide the basis for deciding whether forestry practices are sustainable or not. The incorporation of a hypothetico-deductive framework into management processes is seen as critical for ensuring defensible standards in environmental decision making (Murphy and Noon, 1991; Calver et al., 1998; Lindenmayer, 1999).

The first aim of this paper is to warn against a potential (but often avoidable) problem that may arise during assessments of sustainable forestry objectives within a hypothetico-deductive framework. This problem is the development of monitoring programs that lack sufficient replication, and thus result in statistical analyses with low power. The second aim is to present a step by step guide to experimental planning that will reduce the chance of conducting low power statistical tests. It is suggested that a priori power analysis needs to be an integral part of monitoring program design, and that locally relevant α:β ratios should be determined for each experiment. Although problems related to insufficient power have previously been raised in the context of sustainable forestry (Calver et al., 1998; Burgman and Lindenmayer, 1998; Lindenmayer, 1999), the coverage in these sources lacks detail.

2. Sustainable forest management: definition, monitoring and the detection of change

The ability to assess whether forestry practices are sustainable is, broadly, a three step process. In chronological order, these steps are (a) defining the objectives of sustainable forestry, (b) identifying and monitoring indicators of these objectives, (c) detecting changes in the level of these indicators from a baseline state. The magnitude and direction of change (if it exists) can then be used to assess whether forestry practices are sustainable, and to make appropriate management decisions (Turner and Lambert, 1997).

2.1. *Defining the objectives of sustainable forestry*

Defining the objectives of sustainable forestry is critical for its assessment. Without a clear, unambiguous statement about what sustainable forest management hopes to achieve, no meaningful assessment of forestry practices can be made. At present, a serviceable (and internationally accepted; Brand, 1997) definition of sustainable forestry exists in the form of the criteria outlined in the Santiago Declaration of 1995 (see Turner and Lambert, 1997). These criteria are based upon current social, ecological and economic values.

2.2. *Identification and monitoring of indicators*

Once sustainable forestry has been defined, the next step in the assessment process is to identify indicator variables that provide a measure of sustainable forestry objectives. The identification of indicators (and the subsequent development of methods and protocols for their measurement) is a complex process, and is beyond the scope of this work. Burger and Kelting (1999) provide an example of indicators and measurement methods that have been developed for one common sustainable forestry objective; the maintenance of soil productivity.

After indicators have been selected, data need to be gathered to describe the way they change (or do not change) in time and space. The ability to detect change pre-supposes that the natural or 'sustainable' state of indicator variables can be defined (Burger and Kelting, 1999). Such definitions are critical for the assessment of sustainable forestry objectives. If indicators of sustainable forestry do not have accepted standards against which change can be measured, the sustainability of forestry practices cannot be determined. In order to determine the 'sustainable' levels of indicator variables, data from areas not affected by forestry practices (control sites) need to be collected (Keough and Mapstone, 1995). In addition, it is important to define a critical level of change in the measured variables, i.e., a level of change that is deemed to be necessary to trigger a management response (Keough and Mapstone, 1997).

2.3. Detecting change

Detecting change in indicators of sustainable forestry can be performed with the aid of statistical procedures. Statistical procedures can be used to detect change with a known level of certainty, and thus can greatly aid environmental decision making. In the context of sustainable forest management, if the change in an indicator is greater than a (predetermined) critical level, sustainable forestry objectives are not being met, and forestry practices need to be altered.

In summary, assessments of the sustainability of forestry practices require, a priori, a sound definition of sustainable forestry objectives. Once these objectives have been defined, on ground indicators of these objectives need to be identified. Monitoring of indicators, and the subsequent use of statistical procedures to determine the magnitude of the effect forestry practices have on indicator variables can then be used to make the appropriate management decisions. A diagrammatic summary of the process described above is presented in Fig. 1.

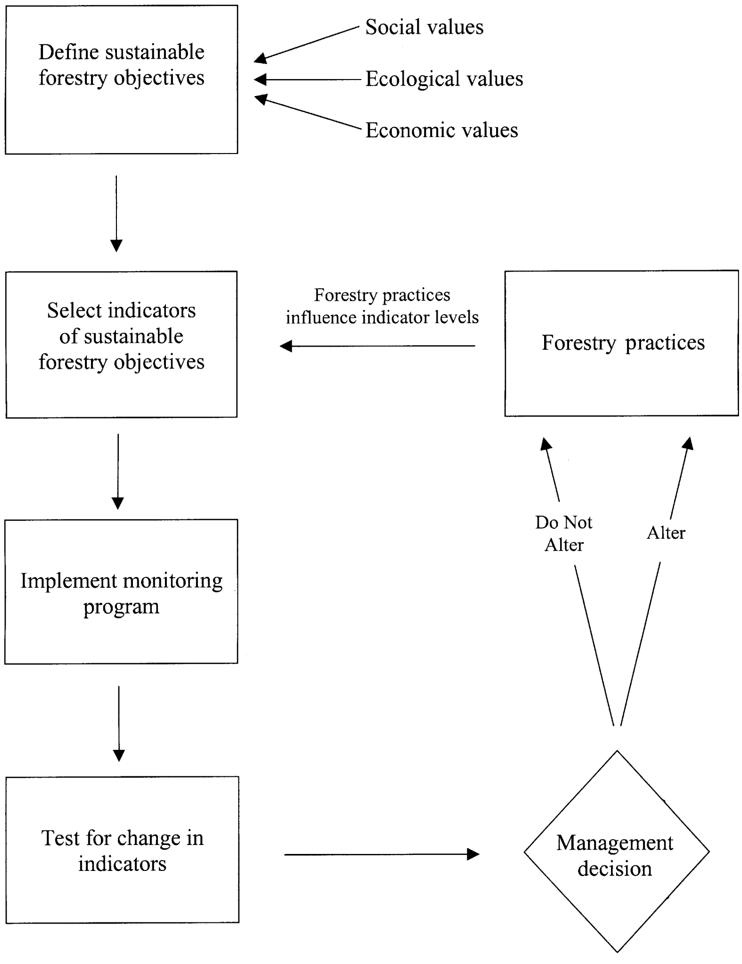

Fig. 1. A process for assessing sustainable forestry objectives.

3. Power analysis and sustainable forest management

3.1. Type I and Type II errors

> . . . all decisions made from monitoring data have the potential to be wrong (Keough and Mapstone, 1993, p. 773).

Because data are often variable, and because statistical decisions are probabilistic, the result of a statistical test may be incorrect (Keough and Mapstone, 1993, 1995). There are two kinds of incorrect decisions. The first (a Type I error) can be made when the result of a statistical test is significant (i.e., when the null hypothesis is rejected). The functional consequence of a Type I error is the rejection of a null hypothesis when it should have been accepted. The second (a Type II error) can be made when the result of a statistical test is not significant (i.e., when the null hypothesis is accepted). The functional consequence of a Type II error is the acceptance of a null hypothesis when it should have been rejected. For any statistical test, the probability of making a Type I error is defined as α and the probability of making a Type II error is defined as β. Both Type I and Type II errors, along with the correct inferences that can be drawn from a statistical test, are shown in Table 1.

In traditional statistical approaches, the probability of making a Type I error (α) is set at 0.05 (Mapstone, 1995). Thus the chance of making a Type I error is small, and significant results can be reported with a high degree of confidence. The probability of making a Type II error (β), however, is usually not controlled (Keough and Mapstone, 1997), and is commonly high or not reported in studies within a number of research disciplines (psychology: Cohen, 1962; Sedlmeier and Gigerenzer, 1989; fisheries and aquatic science: Peterman, 1990; environmental impact studies: Fairweather, 1991; Mapstone, 1995; music research: Daniel, 1993). This is of concern, as it means that the power of statistical procedures is unknown.

3.2. Statistical power

The power of a statistical test is defined as the probability that it will detect a significant difference or relationship between measured variables, if indeed one exists (Cohen, 1988). Power $= 1 - \beta$, and is related to α, the effect size, the sample size and the sample variance. Although the specific form of the relationship between power and these other parameters depends on the statistical method being used, the general relationship can be described (Burgman and Lindenmayer, 1998) by the expression

$$\text{Power} \propto \frac{\text{ES} \times \alpha \times \sqrt{n}}{\sigma} \tag{1}$$

where ES is the effect size, α the Type I error rate, n the sample size and σ is the standard deviation. The form of the relationship is such that power increases as effect size, α and n increase, but decreases as the data become more variable.

A statistical test with low power will result in a high probability of making a Type II error; detecting a non-significant effect when a significant effect does in fact exist. In the context of detecting change in indicators of sustainable forestry, making a Type II error is likely to have important implications.

In the majority of cases, the level of sustainable forestry indicators will change (either increase or decrease) if the objectives of sustainable forestry are not being met. If, e.g., the indicator in question

Table 1
Errors in statistical decision making[a]

Real situation	Conclusion from a statistical test	
	Impact	No impact
Impact	Null hypothesis rejected (correct decision). An impact exists and is detected	Null hypothesis accepted (incorrect decision; Type II error (β)). An impact exists, but is not detected
No impact	Null hypothesis rejected (incorrect decision; Type I error (α)). No impact exists, but one is detected	Null hypothesis rejected (correct decision). No impact exists and no impact is detected

[a] α and β are the probabilities of making Type I and Type II errors, respectively. After Fairweather (1991) and Keough and Mapstone (1995).

is 'soil erosion', unsustainable forestry practices will cause the level of this indicator to rise. A non-significant result from a statistical procedure designed to test the null hypothesis that forestry practices have no effect on the level of soil erosion is likely to be interpreted as conformation that sustainable forestry objectives are being met. If a Type II error has been made, however, the result of the statistical procedure is incorrect; no significant effect has been detected when one does in fact exists. In terms of the above example, forestry practices are found to have no effect on the level of soil erosion when, in reality, they are causing soil erosion to increase. This type of error could have dire environmental consequences (Peterman, 1990; Fairweather, 1991). If Type II errors are being made and incorrect non-significant results are being accepted, forest managers will think that forestry practices are sustainable when in fact they are not.

In order to rectify this potential problem, it is necessary to estimate the probability of making a Type II error before the statistical test is performed. A priori statistical power analysis enables this to be achieved.

For specified levels of α and effect size, and with an estimate of the variance in the underlying population, a priori power analysis can be used to determine how many samples are needed to achieve a predetermined level of power. As power $= 1 - \beta$, a power estimate enables an estimate of β, the probability of making a Type II error. The success of this process requires an estimation of the critical effect size (i.e., a specification of the critical level of change discussed in Section 2.2). Ideally, critical effect size should be estimated on the basis of pre-existing data, ecological theory or legislative requirements. Information from these sources, however, is often lacking and estimations of effect size usually lack a sound basis (Keough and Mapstone, 1997). In situations where information for estimating effect size does not exist, conventions proposed by Cohen (1988, 1992) may be used.

A priori power analysis should be used in the planning stage of an experiment to better understand the chance that an error (either Type I or Type II) will be made for a particular sample size, effect size and variance. A priori power analysis, however, only approximates the realised error rates (due, in most situations, to inaccuracies in the estimate of population variance; Keough and Mapstone, 1997). Consequently, an additional power analysis (often called post hoc or retrospective power analysis) should be performed in order to determine the realised level of β. This procedure is only necessary if the result of the statistical test is not significant; a significant result means, by definition, that power has been sufficient to detect the critical effect (Fairweather, 1991). In the case of a non-significant result, calculation of the realised value of β (the probability of making a Type II error) will inform decision makers of the likelihood that the result of the statistical test is incorrect. If the risk of making a Type II error is known, forest managers can interpret the result with a known level of confidence. This information provides a sound basis from which forest managers can decide whether the objectives of sustainable forestry are being met.

4. Liberating α in the power analysis process

Underwood (1997, p. 91) 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.

If setting the probability of making a Type I error at 0.05 is so arbitrary, why should this convention be retained? There are strong arguments that it should not.

4.1. Problems with traditional decision criteria

Traditionally, the probability of making a Type I error (α) when performing a statistical procedure is set at 0.05 (Mapstone, 1995). While the probability of making a Type II error (β) may sometimes be correspondingly small, in most situations it is not (e.g., Cohen, 1988). In general, the relationship between α, β, effect size, sample size and sample variance (see Eq. (1)) means that when α is 0.05, sample size or effect size (or both) must be very large for β to be correspondingly small (Mapstone, 1995). Thus the functional consequence of setting α at 0.05 is that β is usually much larger. As was discussed in Section 3.2, this may have dire environmental consequences.

In addition, when α is smaller than β there is an implicit assumption that making a Type I error is more important than making a Type II error. For example, in a case where $\alpha = 0.05$ and $\beta = 0.20$ (i.e., power $=$ 0.80; a value that is often considered sufficient), it is assumed that making a Type I error is four times more serious than making a Type II error (Cohen, 1992). In environmental impact studies, the consequences of making a Type II error (i.e., concluding that there is no anthropogenic effect when in fact there is) may be much more serious than the consequences of making a Type I error (Peterman, 1990; Fairweather, 1991). In other situations, however, the consequences of making a Type I error may indeed be more important. What is needed, then, is a flexible protocol which enables researchers and other stakeholders to decide, for a particular set of circumstances, which kind of error is more important. For statistical purposes, this decision can be expressed as a ratio of α:β.

4.2. Determining the ratio of α:β

Decisions about the appropriate ratio of α:β should be made before any data are collected, and should be based on the relative cost of making each kind of error. For example, if the cost of making Type I and Type II errors is considered equal, the ratio of α:β should be set at 1:1. If, however, the cost of making a Type II error is considered twice as important as making a Type I error, the ratio of α:β should be set at 2:1. It is important to remember that the costs of statistical errors may be biological, social or economic, and all of these fields must be considered when determining an appropriate α:β ratio.

Determining a locally relevant α:β ratio can be incorporated into a priori power analysis. Once the ratio of α:β has been determined, however, it remains fixed throughout the rest of the process. The actual values of α and β for a given experiment are influenced by the degree of statistical risk that is deemed acceptable and the resources (time, money, etc.) that are available. Consequently, the actual values of α and β may change a number of times during the planning phase of an experiment. This will be exemplified in Section 5.2.

Liberating α and setting the ratio of α:β to reflect site-specific social, economic and biological circumstances is a logical approach to testing the objectives of sustainable forestry. Determination of an appropriate α:β ratio before data are collected will reduce the chance of making a statistical error (whether it be Type I or Type II) that will have serious negative implications.

5. A step by step guide to experimental planning and power analysis

This section describes six steps involved in planning an experiment using a priori power analysis and incorporating the determination of an appropriate α:β ratio. The steps involved in this process are based on the work of Keough and Mapstone (1993, 1995, 1997) and Mapstone (1995). The theoretical step by step guide is followed by a simple hypothetical example to show how the process would work in the field.

5.1. Six steps to experimental planning

Step 1. Identify an experimental design that will provide an adequate test of the null hypothesis in question. For testing the null hypothesis of no impact in environmental impact studies, an MBACI (Multiple Before–After, Control–Impact) design is recommended (Keough and Mapstone, 1995, 1997; Burgman et al., 1998).

Step 2. Select indicator variables for monitoring. Some indicators that can be used for testing sustainable forestry objectives have already been developed (e.g., McLaren et al., 1998; Burger and Kelting, 1999; Rab, 1999).

Step 3. Step 3 is divided into three sub-steps, (a), (b) and (c). All three sub-steps must occur after Step 2, but the particular order is not important.

(a) Select a critical effect size, i.e., the smallest effect that it is important to detect. When assessing the objectives of sustainable forestry, the critical effect size should be the smallest change in an indicator variable that is deemed necessary to stimulate management action.

(b) Estimate the variance of the indicator variable in the underlying population. Variance estimates are often difficult, but can be based on past studies or pilot data. When such data are available, an estimate of the population standard deviation (the value required by power analysis software) for *t*-tests and simple

ANOVAs can be made using the formula

$$\text{Population S.D.} = \sqrt{\frac{\sum(x - \bar{x}_g)^2}{n - k}} \quad (2)$$

where x is the value of the measured variable, $\bar{x}_g$ the group mean, n the total sample size across groups and k is the number of groups. This formula calculates the square root of the pooled within-groups variances and is the best way to estimate the population standard deviation from pre-existing data (E. Erdfelder, personal communication).

(c) Establish the relative cost of making Type I and Type II errors and use this information to set the ratio of α:β. The final decision about the relative cost of each type of error may be based on information from social, economic or biological sources. It is important to note that the relative cost of the two types of error and the subsequent ratio of α:β will vary from study to study (Keough and Mapstone, 1997).

Step 4. Remembering that the ratio of α:β is already fixed, determine the largest levels of α and β that are acceptable. If, for example, the consequences of making a Type II error are considered serious, the chance of making a Type II error may be initially set at 0.05 (i.e., a 5% change of making a Type II error is considered acceptable — power = 0.95). If (in Step 3(b)) the ratio α:β has been defined as, say, 2:1, the level of α would then be set at 0.1. These initial levels of α and β may be referred to as desired or ideal values and may need to be increased later in the process (Mapstone, 1995). In the context of sustainable forestry, Step 4 should involve discussion between forest managers, scientists and other interested parties.

Step 5. Use a priori power analysis techniques to determine the level of replication required to detect the critical effect size at the predetermined levels of α and β. Use this information to estimate the cost of the monitoring program that will be necessary, and determine its feasibility. If the cost of the monitoring program is within acceptable limits, proceed with data collection. If the cost of monitoring is considered too great, there are three options: (a) accept higher error rates (higher levels for α and β), (b) find cheaper monitoring techniques or (c) select new variables that will be less time consuming and/or cheaper to measure. Option (a), (b) or (c) (or a combination of the three) should be used to generate a satisfactory compromise between the risk of making statistical errors and the cost of the monitoring program. It is important to note that none of these options affect the ratio of α to β. If the levels of α and β rise, they rise proportionately, thus there is no bias in the type of statistical error that is made (Keough and Mapstone, 1993).

Step 6. Once data have been collected, proceed with the test of the null hypothesis. Use the final value of α as the criterion for accepting or rejecting the null hypothesis of no effect. If the statistical procedure indicates a significant effect, this result may be accepted with a known degree of confidence. If the result indicates a non-significant effect, use post hoc power analysis to calculate the realised level of β. This is necessary because the level of β defined by a priori power analysis is just an estimate and may change when data for the planned study are collected. The realised level of β determined by post hoc power analysis provides researchers with the actual power of their test and the probability that a Type II error has been made. This information can then be used to assess the risk associated with accepting the non-significant result. Steps 1 through 6 are shown diagrammatically in Fig. 2.

5.2. Power analysis and sustainable forest management: an example

The experimental planning process described above is highly relevant to experiments designed to test the objectives of sustainable forestry. While the broad goals of sustainable forestry are unlikely to change, particular objectives may vary from country to country and from region to region depending on particular social, economic and ecological pressures (Brand, 1997). Liberating α and enabling the α:β ratio to be based on the relative costs of Type I and Type II errors, and then balancing error risk against monitoring program feasibility provides a flexible protocol that can be fitted to any set of circumstances. A hypothetical example of the flexibility of this system, and its consequent utility for assessing sustainable forestry objectives, is provided.

Background. An international conference about sustainable forest management has just been held. The conference has drawn substantial media attention and a particular focus of the press has been the danger arboreal marsupials face from logging. Local news

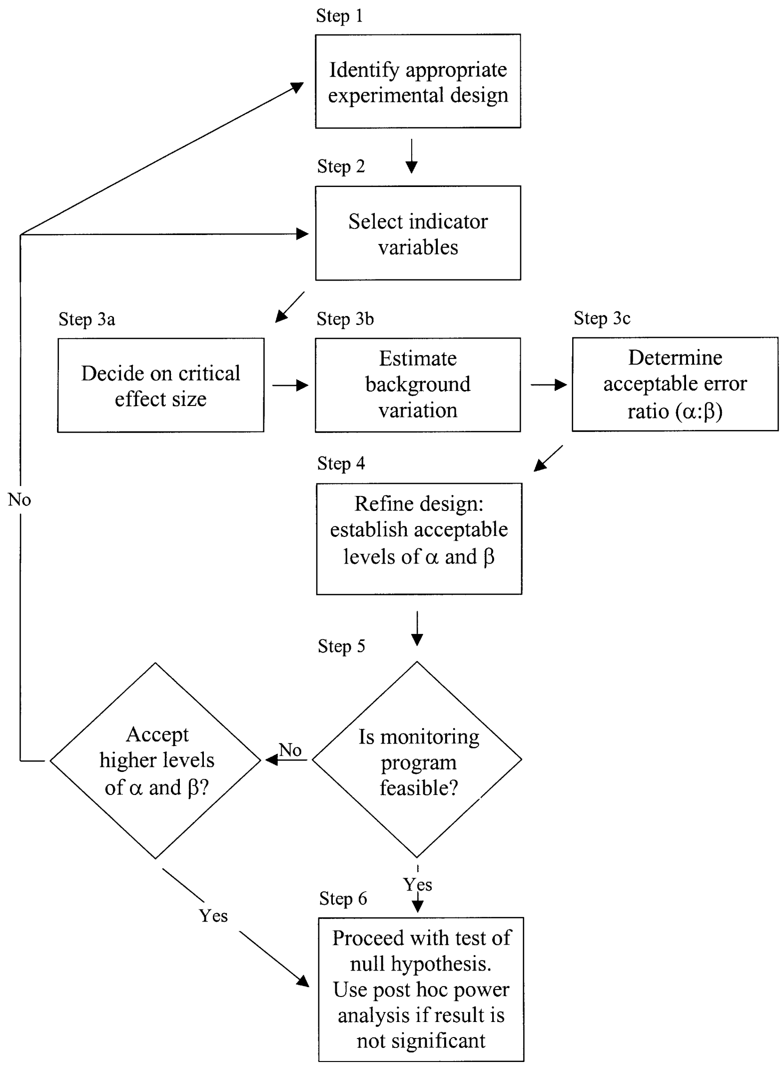

Fig. 2. A six step approach to experimental planning using a priori power analysis. Based on a diagram in Keough and Mapstone (1997).

bulletins have been presenting file tape of cute, frightened-looking possums juxtaposed against loud timber harvesting machinery dragging felled trees across a devastated landscape. The issue has caught the public's attention and anti-logging protesters are rapidly gaining support.

Consequently, the local land management authority has decided to fund two experiments to test the sustainability of timber harvesting practices in the region concerned. One experiment is planned for Area A and the other for Area B. Both these areas are in the same geographical region and are dominated by lowland mixed species forest. In general, the experiments are designed to test the international sustainable forestry objective of biological diversity (Turner and Lambert, 1997), but a more important local objective is to see whether timber harvesting is reducing the abundance of arboreal marsupials. Specifically, the experiments are designed to test the null hypothesis that timber harvesting has no effect on the abundance of large hollow trees. Large hollow trees was chosen as a surrogate measure for arboreal marsupials, as it is known this resource is important for arboreal marsupial populations (e.g., Lindenmayer et al., 1990, 1991;

Burgman and Lindenmayer, 1998). The plan is to test whether there is any difference in the number of large hollow trees between mature forest stands and regrowth stands. A simple two sided *t*-test is chosen as the appropriate statistical technique.

It soon becomes obvious to the research team that the social, economic and ecological situation differs markedly between Areas A and B.

Area A. Area A is relatively affluent and derives the majority of its economic turnover from three industries: tourism, dairy farming and timber harvesting. The core of the tourism industry is associated with night-time forest walks where arboreal fauna are the main attraction. Consequently, the persistence or otherwise of local arboreal marsupials will have significant economic implications.

Area B. In contrast to Area A, timber harvesting is by far the largest industry in Area B. In a social climate of high unemployment and few alternative prospects, the timber industry provides jobs for 60% of the local population. The forests of Area B contain arboreal marsupials, but they are not considered to be of local economic importance.

Planning the experiment. The research team decides to use the approach to experimental planning outlined in Section 5.1. They believe this decision is justified because of the different set of circumstances apparent in Areas A and B. Step 1 (defining a design) and Step 2 (choosing indicator variables) have already been taken, so the research team moves onto Step 3(a), determining a critical effect size. It is decided that critical effect size calculations should be based on available data about the requirements of five hollow dwelling species (the common brushtail possum (*Trichosurus vulpecula* Kerr), the eastern pygmy-possum (*Cercartetus nanus* Desmarest), the yellow-bellied glider (*Petaurus australis* Shaw), the sugar glider (*P. breviceps* Waterhouse) and the common ringtail possum (*Pseudocheirus peregrinus* Boddaert) that are known to exist in the forests of Areas A and B. Ranges for the number of nest trees used per hectare for these species are presented in Gibbons and Lindenmayer's (1997) Table 3. The median of each range was selected, and then the mean of these values (14.7) was used to calculate the critical effect size for Area A.

Existing data from lowland mixed species forest of southwestern Victoria (Nelson, unpublished data) indicate that the average number of large hollow trees per hectare in mature forest stands is 18.9. The difference between 18.9 and 14.7 represents a 22% reduction in the number of hollow trees per hectare, and this value is used as the critical effect size in Area A. In Area B, however, local community groups argue that retaining 14.7 large hollow trees per hectare is unnecessary: after all, this is much higher than the value defined by current harvesting prescriptions (Gibbons and Lindenmayer, 1997). Consequently, it is agreed to reduce this value by 25 to 11.0. Thus the difference between 18.9 and 11.0 (about 42%) becomes the critical effect size for Area B.

Computer programs that conduct a priori power analysis require the user to input two means (termed Mean 1 (μ_1) and Mean 2 (μ_2)). In Area A, μ_1 refers to the estimated existing number of large hollow trees per hectare in mature forest stands (18.9) and μ_2 refers to the number of large hollow trees per hectare that has been accepted as a reasonable minimum value (14.7). In Area B, μ_1 and μ_2 are 18.9 and 11.0, respectively.

Next, the same data that was used to estimate the number of large hollow trees in mature forest stands (Nelson, unpublished data) is used to estimate the variation in the underlying population for the indicator variable in question (Step 3(b)). These data, based on 52 one hectare sites, estimate the population standard deviation to be 10.22. This information, along with the μ_1 and μ_2 values, will be used to calculate the effect size indexes for Areas A and B as part of the a priori power analysis (see Table 2).

Step 3(c) involves setting an appropriate α:β ratio. After discussion with stakeholders in Area A, the research team decides that the costs of making a Type II error (failing to detect an effect when one exists) is twice as serious as making a Type I error (detecting an effect when one does not exist). The cost of making a Type II error is deemed to be serious because it could result in a reduction in local arboreal marsupial populations and a subsequent loss of tourism dollars. Thus making a Type II error could lead to both ecological and economic costs. In contrast, making a Type I error could result in an unnecessary reduction in timber harvesting, an action that would not have any adverse ecological impacts and is predicted to have only a moderately negative economic effect. Consequently, the ratio of α:β is set at 2:1.

Because the cost of making a Type II error is so substantial, the research team (in conjunction with

Table 2
Details of the a priori power analysis conducted for Areas A and B[a]

	Area A, Analysis 1	Area A, Analysis 2	Area B
α:β	2:1	2:1	1:2
α	0.10	0.20	0.05
Estimated μ_1 (mature stands)	18.9	18.9	18.9
Estimated μ_2 (regrowth stands)	14.7	14.7	11.0
Critical effect size	22%	22%	42%
Estimated σ	10.22	10.22	10.22
Effect size index (δ), $\delta = \lvert\mu_1 - \mu_2\rvert/\sigma$[b]	0.411	0.411	0.773
Power	95%	90%	90%
Samples per group	129	79	37

[a] Power was calculated using the software package GPOWER (Faul and Erdfelder, 1992).
[b] The formula for δ is different for different statistical tests. See Cohen (1992).

other interested parties) decide that that the risk of making a Type II error should be no greater than 5%. Thus the ideal value for β is set at 0.05 (power $= 0.95$) and, as α:β equals 2:1, α becomes 0.10. The information collected so far is exposed to a priori power analysis (Analysis 1 in Table 2) and the results show that the sampling effort required to achieve these levels of α and β are very large; a total of 258 sites will be required. The resources to monitor 258 sites are not available, so the research team, along with other interested parties, meet and reassess the situation. After due consideration, it is decided that a 10% risk of making a Type II error will be accepted. Thus new levels of α and β are set at 0.20 and 0.10, respectively (the original 2:1 ratio does not change). This new information is expose to a priori power analysis (Analysis 2 in Table 2) and the results show that a total of 158 sites will be required. Monitoring 158 sites will be expensive, but both the research team and members of the local community are not prepared to increase the risk of making a Type II error. As the research team does not believe that there is any other way of reducing monitoring costs, and because everyone involved in the process recognises the importance of the issues at hand, the local land management authority agrees to the cost of the project. The actions just described correspond to Steps 4 and 5 of the study planning process.

In Area B, it is deemed that making a Type I error will be twice as serious as making a Type II error. The major cost of making a Type I error would be an unnecessary reduction in timber harvesting operations, an outcome that would adversely affecting the local human population who depend so heavily on the timber industry. The cost of making a Type II error would be a reduction in the distribution and abundance of arboreal marsupials, but this is considered less important than the potential adverse effects faced by the local human inhabitants. Consequently, the ratio of α:β is set at 1:2.

From this point, the process that is followed is the same as that described for Area A. It is decided that a 5% risk of making a Type I error is acceptable, and thus the ideal values of α and β are set at 0.05 and 0.10 (power $= 0.90$). The results of a priori power analysis (Table 2) indicate that 74 sites are required. This level of monitoring is considered acceptable by the local land management authority. In this case, the number of samples required is reduced due to the larger critical effect size specified for Area B. The relationship between power and sample size for each of the three analyses in Table 2 is shown graphically in Fig. 3.

The next step (Step 6) taken by the research team is to establish sites, collect real data and use the predefined statistical technique to test the null hypothesis of no effect. Hypothetical data have been used to generate means, standard deviations, realised effect sizes and P-values for Areas A and B. These variables are shown in Table 3.

The result of the statistical test for Area A indicates that there is a significant difference in the number of large hollow trees between mature and regrowth sites. Examination of the realised effect size (19.3%; Table 3) shows that this is a meaningful result in the context of the present study where a 22% reduction was considered important. While larger than the tradi-

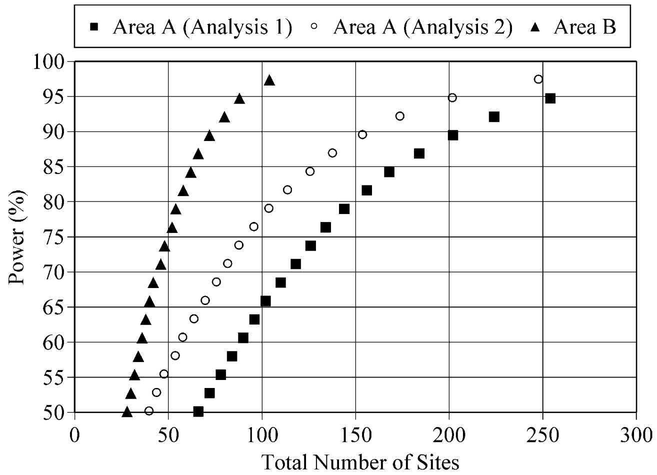

Fig. 3. The relationship between power and the number of sites for Area A (Analyses 1 and 2) and Area B. This analysis assumes that there are an equal sample size (total number of sites) in each forest type (mature and regrowth).

tional cut off point of 0.05, the realised level of α (0.088; Table 3) is considered significant in the context of Area A where α has been defined as 0.20.

The result of the statistical test for Area B indicates that there is no difference in the number of large hollow trees between mature and regrowth sites. As recommended in Step 6 (Section 5.1), the research team conducts post hoc power analysis to determine the realised level of β, and hence the exact power of their test. Post hoc power analysis using the realised effect size of 24.5% (Table 3) reveals that power is only 0.34, well below the agreed upon value of 0.90. However, post hoc power analysis is only meaningful if the critical effect size is used, as this is the difference between means that it is important to detect (Thomas, 1997). When the critical effect size for Area B (42%) is used in the post hoc power analysis, power is 0.75, still well below the agreed upon value of 0.90. This is because the real variability in the data (Table 3) is greater than the estimate of variability used in the a priori power analysis. Further, power analysis using the real level of variability indicates that data from 38 additional sites in Area B are required to achieve a statistical result with a power of 0.90. At this point the research team can either accept the original (non-significant) result with an understanding that it has a 25% chance of being wrong, or they can attempt to secure funding to collect the extra data. Whatever the decision, the research results can be presented with a known degree of confidence, and knowing the accuracy of

Table 3
Means, standard deviations, realised effect sizes and P-values for Areas A and B[a]

	Sample mean	Sample standard deviation	Realised effect size	P-value (realised α)
Area A				
Mature stand	19.12	13.89	19.3%	0.088
Regrowth stand	15.43	13.10		
Area B				
Mature stand	17.66	12.17[b]	24.5%	0.123
Regrowth stand	13.34	11.64[b]		

[a] The data set is hypothetical.

[b] The value for the population standard deviation used in the post hoc power analysis for Area B was 11.91 (see Step 3(b) in Section 5.1 for an explanation of how to calculate this value).

research results provides a basis for superior management decisions.

Although the experiments described above are hypothetical and rather simple (there are, for example, many factors affecting the use of hollows by arboreal marsupials (Gibbons and Lindenmayer, 1997) that are not considered), they touch on some of the challenges that are likely to arise when testing the objectives of sustainable forestry. There is little doubt that regardless of the indicators that are used or the hypotheses that are tested, different forest management areas may have social, ecological or economic influences that necessitate alternative assessments of statistical risk. The above examples shows how a process of study planning incorporating a priori power analysis and flexible α:β ratios is helpful for testing the objectives of sustainable forestry in an heterogeneous world.

6. Conclusion

The concept of sustainable forestry is gathering attention, and sustainable forestry objectives have been defined on an international scale. At local levels, work has begun to develop indicator variables that can be used to assess whether the objectives of sustainable forestry are being met. It is widely accepted that such assessment should proceed within a hypothetico-deductive framework; i.e., using statistical procedures to test null hypotheses of no effect.

The results of statistical procedures, however, are probabilistic; we can never be sure that they are correct. Before statistics can be used to aid environmental decision-making, the risk of committing a statistical error must be determined. Although traditional statistical approaches provide information about the risk of making a Type I error, the risk of making a Type II error is often unknown. A priori power analysis, however, can be used to estimate the risk of making a Type II error before data have been collected. Using this process in the planning phase of an experiment can severely reduce the chance that a Type II error will occur. In the event of a non-significant result, post hoc power analysis can be used to determine the probability that a Type II error has occurred. In the context of assessing sustainable forestry objectives, the use of both a priori and post hoc power analysis can help forest managers make better management decisions. There is little doubt that the use of power analysis in the planning stage of experiments designed to test sustainable forestry objectives will improve the quality of sustainable forest management.

It has also been suggested that statistical assessments of sustainable forestry can be more relevant if α is liberated from its traditional value of 0.05. One consequence of liberating α is that statistical procedures can incorporate local biological, social and economic influences into the analysis. This is highly relevant for assessments of sustainable forestry, as different forest management units may have markedly different biological, social and economic objectives. It is hoped that in all fields of experimental science, developing a locally meaningful α:β ratio prior to data collection will replace the traditional, arbitrary and sometimes detrimental convention of setting α at 0.05.

Acknowledgements

Rob Campbell and Kylie McKenzie read earlier drafts of this manuscript and contributed to its clarity. Jenny Nelson provided the data upon which the power analysis in the hypothetical example is based and also made suggestions that improved the reality of the example. Edgar Erdfelder provided information that helped to clarify the way population standard deviation can be estimated when *t*-tests and simple ANOVAs are being used.

References

Brand, D.G., 1997. Criteria and indicators for the conservation and sustainable management of forests: progress to date and future directions. Biomass Bioenerg. 13, 247–253.

Burger, J.A., Kelting, D.L., 1999. Using soil quality indicators to assess forest stand management. For. Ecol. Mgmt. 122, 155–166.

Burgman, M.A., Lindenmayer, D.B. 1998. Conservation Biology for the Australian Environment. Surrey Beatty, Chipping Norton, Sydney, NSW.

Burgman, M., Ades, P., Hickey, J., Williams, M., Davies, C., Maillardet, R., 1998. Methodological guidelines for the utility and statistical validity of survey and monitoring programs. Unpublished Report to the Forest and Wood Products Research and Development Corporation.

Calver, M.C., Dickman, C.R., Feller, M.C., Hobbs, R.J., Horwitz, P., Recher, H.F., Wardell-Johnson, G., 1998. Towards resolving conflict between forestry and conservation in Western Australia. Aust. For. 61, 258–266.

Cohen, J., 1962. The statistical power of abnormal-social psychological research: a review. J. Abnorm. Soc. Psychol. 65, 145–153.

Cohen, J., 1988. Statistical Power Analysis for the Behavioural Sciences, 2nd Edition. Lawrence Erlbaum, Hillsdale, NJ.

Cohen, J., 1992. A power primer. Psychol. Bull. 112, 155–159.

Commonwealth of Australia, 1998. A Framework of Regional (Sub-national) Level Criteria and Indicators of Sustainable Forest Management in Australia. Pirie Printers, Canberra, ACT.

Daniel, T.D., 1993. A statistical power analysis of the quantitative techniques used in the Journal of Research in Music Education, 1987–1991. Ph.D. Thesis. Auburn University, Alabama.

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

Faul, F., Erdfelder, E., 1992. GPOWER: a priori, post-hoc, and compromise power analyses for MS-DOS (computer program). Department of Psychology, Bonn University, Bonn.

Gibbons, F., Lindenmayer, D.B., 1997. Developing tree retention strategies for hollow-dependent arboreal marsupials in the wood production eucalypt forests of eastern Australia. Aust. For. 60, 29–45.

Keough, M., Mapstone, B., 1993. Protocols for designing marine monitoring programs associated with BEKM operations. In: Proceedings of the 47th Appita Annual General Conference. APPITA, Melbourne, Vic., pp. 773–778.

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

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

Lindenmayer, D.B., 1999. Future directions for biodiversity conservation in managed forests: indicator species, impact studies and monitoring programs. For. Ecol. Mgmt. 115, 277–287.

Lindenmayer, D.B., Cunningham, R.B., Tanton, M.T., Smith, A.P., Nix, H.A., 1990. Habitat requirements of the mountain brushtail possum and the greater glider in the montane ash-type eucalypt forests of the Central Highlands of Victoria. Aust. Wildl. Res. 17, 467–478.

Lindenmayer, D.B., Cunningham, R.B., Tanton, M.T., Nix, H.A., Smith, A.P., 1991. The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, southeast Australia: III. The habitat requirements of Leadbeater's possum *Gymnobelideus leadbeateri* and models of the diversity and abundance of arboreal marsupials. Biol. Conserv. 56, 295–315.

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

McLaren, M.A., Thompson, I.D., Baker, J.A., 1998. Selection of vertebrate wildlife indicators for monitoring sustainable forest management in Ontario. For. Chronol. 74, 241–248.

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

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

Platt, J.R., 1964. Strong inference. Science 146, 347–353.

Popper, K.R., 1968. The Logic of Scientific Discovery. Hutchinson, London.

Rab, M.A., 1999. Measures and operating standards for assessing Montreal process soil sustainability indicators with reference to Victorian Central Highlands forest, southeastern Australia. For. Ecol. Mgmt. 117, 53–73.

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

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

Turner, J., Lambert, M., 1997. Development of indicators of sustainable management in Australian forests. In: Hale, P., Lamb, D. (Eds.), Conservation Outside Nature Reserves. Centre for Conservation Biology, University of Queensland, pp. 471–475.

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, Cambridge, UK.