

Provided for non-commercial research and educational use only.
Not for reproduction or distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Use of confidence intervals to demonstrate performance against forest management standards

Terry Walshe^{*}, Brendan Wintle, Fiona Fidler, Mark Burgman

School of Botany, University of Melbourne, Vic. 3010, Australia

Received 1 December 2006; received in revised form 18 April 2007; accepted 30 April 2007

Abstract

The objective of continuous improvement embedded in forest management standards relies on the capacity of management to respond appropriately to evidence of performance provided by monitoring. This evidence is rarely unequivocal. Under a null hypothesis of no effect, two kinds of errors in interpretation are possible—inferring an effect where none exists (Type I error) and inferring no effect when in fact one exists (Type II error). If the monitoring relates to possible improvement in growth or yield then a Type I error leads to false optimism and a Type II error to false pessimism. If monitoring concerns a potential environmental or social impact, a Type I error implies alarmism and a Type II error a false sense of security.

Explicit consideration of statistical power in designing and interpreting monitoring data is an effective buffer against these errors. However, strict application of statistical power may be impractical. In particular, the requirement to specify tolerable error rates and effect sizes will be difficult in many circumstances where the perspectives of managers, auditors or stakeholders are contested or perceived to be arbitrary or vague. We advocate the use of confidence intervals as an alternative to power calculations. Confidence intervals offer an accessible approach to communicating performance under a standard and the extent to which a monitoring program is able to distinguish compliance from non-compliance. We illustrate these arguments and tools through a hypothetical example involving a proposed change in silviculture where the magnitude of gains in yield and environmental impacts are unclear.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Forest management standards; Monitoring; Performance communication; Type I and Type II error; Statistical power; Confidence intervals

1. Introduction

Forestry standards seek to synthesize what is understood to be best practice. Examples include the Forest Stewardship Council International Standard (FSC, 2000), criteria and indicators developed under the Montreal Process (1999), and the Australian Forestry Standard (Standards Australia, 2003). Implicitly, accreditation under a standard asserts a company's commitment to best management and continuous improvement. The validity of this assertion rests on demonstrated compliance. This paper outlines how data gathered in monitoring can be analysed and communicated in a way that is accessible to managers, auditors and stakeholders. Its motivation is facilitation of evidence-based continuous improvement.

The criteria and indicators contained in a standard represent an attempt to encapsulate values associated with forests in a form that is amenable to measurement. Suter (1993) describes a hierarchical process for translating broad management goals into measurement endpoints. Management goals are statements that embody broad objectives. They are often ambiguous or vague, but carry with them a clear social or organizational mandate. Measurement endpoints are elements that can actually be measured. They can be regarded as operational definitions of management goals. In their emphasis on measurable outcomes as a means of demonstrating compliance, forestry standards inevitably tend toward reductionism at the expense of holistic ecosystem-level perspectives on forest processes. We regard the specification of measurable endpoints as a necessity in management systems where the insights from monitoring underpin adaptive management and continuous improvement.

The technical task of gathering and interpreting monitoring data within a continuous improvement framework takes place

^{*} Corresponding author. Tel.: +61 3 8344 3201; fax: +61 3 9348 1620.
E-mail address: twalshe@unimelb.edu.au (T. Walshe).

against a complex social and political background involving conflicting values and multiple narratives regarding the magnitude and seriousness of various impacts (Raison et al., 2001). Managers may be over-confident in the effectiveness of their actions (Morgan and Henrion, 1990; Ludwig et al., 1993) and tend to regard the claims of environmentalists and others as alarmist. The attitude of community stakeholders toward natural resource managers is often characterized by skepticism rather than trust (Bocking, 2004). They may be inclined to view the claims of managers as imbuing a false sense of security. Hard data and scientific rigor are often invoked as the ultimate arbiters of contested claims. However, there is a distinct tendency for people to draw firm conclusions from meager data that are inconsistent with what they themselves might regard as a reasonable burden of proof (Tversky and Kahneman, 1971).

Statistical analysis provides a basis for assessing the extent to which data is consistent with alternative assertions. There are several approaches including frequentist and Bayesian analysis. The most commonly-used and taught approach is frequentist analysis involving null hypothesis testing, whereby the calculated probability (p -value) of obtaining the observed data given the null hypothesis is based on the same sampling procedure being implemented many times. Our focus here is on frequentist interpretation of data, but we note that Bayesian analyses may be equally appropriate. Bayesian methods combine prior information and the probabilities of obtaining the data under alternative hypotheses to obtain updated estimates of the evidence in favor of the hypotheses. The Bayesian analogue of confidence intervals is Bayesian credible intervals. Quinn and Keough (2002) and McCarthy (2007) provide further discussion of the different schools of statistical inference and their application to the biological sciences.

The standard frequentist approach to the evidence provided by data is poorly placed to deal with typical natural resource management issues (Burgman, 2005). Science has an asymmetric view of evidence as a consequence of its focus on the accumulation of knowledge. It is strongly averse to the possibility of concluding from a study that an effect exists when in fact it does not. Science is less concerned with the possibility of concluding no effect when one in fact exists. This is problematic because forest managers, auditors and stakeholders need to be aware of the possibility of both kinds of errors.

Specifically: Inferences drawn from monitoring data can make two kinds of mistakes (Table 1)—inferring an effect or impact when there is none (Type I error; denoted α) or inferring

no effect or impact when there is one (Type II error; denoted β). Where monitoring involves testing economic aspects of sustainability's triple bottom line (e.g. the effectiveness of a thinning regime or fertiliser treatment on yield), Type I and Type II errors generally imply false optimism and false pessimism, respectively. The costs of wrong inferences are borne directly by the company. Where the issue involves impacts on the environmental or social dimensions of sustainability, a Type I error translates to false alarmism and a Type II error promotes a false sense of security. The costs of Type I errors in these circumstances are again borne largely by the company through (unreasonable) loss of community and/or commercial reputation. The costs of Type II errors are borne by the broader community through (unacknowledged) attrition of public good values or resources.

Standard statistical conventions commonly used in scientific research specify tolerable Type I error rates (commonly $\alpha = 0.05$) but are blind to Type II errors. Statistical power is a measure of the confidence with which we would have detected a particular effect, if one existed (Fowler et al., 1998). It is defined as $1 - \beta$ (i.e. the complement of the Type II error rate). Explicit consideration of statistical power is required when decisions are sensitive to Type II errors as well as Type I errors. The likelihood of Type I and Type II errors in any study or monitoring program will decrease as data accumulate and the reliability of evidence improves.

In consultation with auditors and stakeholders, a monitoring program could consider acceptable thresholds for committing Type I and Type II errors, and the program designed accordingly (Mapstone, 1995; Di Stefano, 2001; Foster, 2001; Field et al., 2004). But what is a tolerable impact? What is a reasonable burden of proof in demonstrating compliance (or non-compliance)? To what extent should the burden of proof be conditioned by the cost of collecting data? If answers to these questions can be given, it is a reasonably straight-forward technical exercise for a statistician to calculate the sampling effort (and the monitoring budget) required to make an assessment of whether or not a forest manager is complying with a standard or criterion.

Strict approaches to the use of statistical power in designing monitoring programs assume that clear answers to these questions are available. But in practice, such clarity rarely exists. Notions of burden of proof and what might be regarded a tolerable impact are not solely technical questions. They involve resolution of the individual and collective judgments and preferences of industry, auditors and stakeholders.

Statistical power is an essential concept in the design of monitoring programs and the interpretation of results. However, our guess is that forest managers, auditors and stakeholders will not immediately resolve the issues and values associated with alternative perspectives on how large an impact is tolerable or what burden of proof should apply. Here we advocate a loose interpretation of statistical power that encourages progressive exploration of these themes using confidence intervals. Confidence intervals provide a simple graphical basis for informed discussion on what might constitute compliance and non-compliance and on the extent to which a monitoring

Table 1
Correct and incorrect inferences from monitoring

	Inference	
	Important effect	Unimportant effect
Actual state		
Important effect	Correct	False negative (Type II error, β)
Unimportant effect	False positive (Type I error, α)	Correct

program is able to distinguish compliance from non-compliance.

The focus of this paper is the statistical treatment and presentation of data gathered through monitoring. Its particular emphasis is sampling error. There are other aspects to designing monitoring programs that are beyond the scope of this paper. We treat neither the broad principles of sampling design (see Philip, 1994), nor details such as selecting reference or control sites (see ANZECC/ARCMANZ, 2000) or measurable indicators (see Prabhu et al., 2001).

The next section explores the concept of statistical power in some detail. An example is used to show how statistical inference can be misused, and how consideration of power insulates against misuse. We then make a distinction between formal and loose interpretations of statistical power and recommend confidence intervals to communicate performance. The discussion emphasizes that blanket prescriptions for demonstrating compliance (or non-compliance) are unlikely to be efficient or operationally feasible. We urge auditors, managers and stakeholders to differentiate aspects or criteria within a standard that are of greater importance from those of lesser importance (recognising the management context of individual circumstances), and to allocate their monitoring resources accordingly.

2. Statistical power

Uncertainty is inevitable in natural resource management, where sampling is undertaken in environments characterized by high variability. Through use of a detailed example, we explore the central importance of uncertainty and statistical power in monitoring. We then outline how confidence intervals can be used as an effective analytical approach to summarize data and also as an effective and accessible means of communicating performance to auditors and stakeholders.

2.1. Context—a hypothetical scenario

Consider the following hypothetical scenario: A company's inventory indicates that plantations established on a particular soil type grow notably slower than other soils. Based on silvicultural research elsewhere, it is proposed that yields may be improved by intense mechanical disturbance pre-establishment. The company is keen to adopt the proposal but encounters opposition from community stakeholders. They claim mechanical disturbance will lead to soil loss and increased stream turbidity which in turn will threaten a rare frog known to occur on the company's estate.

After consultation with stakeholders the company decides to trial the new silvicultural treatment in a subset of catchments to better understand (a) any improvement in yield, and (b) potential impacts on local frog populations. Past inventory records show that without intense mechanical disturbance the average mean annual increment (MAI) 3 years post-establishment is $4.1 \text{ m}^3 \text{ ha}^{-1} \text{ year}^{-1}$. Past turbidity measurements in streams adjacent to recently cut coupes varied around 9.2 turbidity units. The trial involves 3 years of

collecting monitoring data across several coupes and adjacent streams.

Let us say the company collects data from 50 inventory plots in trial coupes and makes 25 observations of stream turbidity. And let us say the mean MAI from the inventory plots was $4.6 \text{ m}^3 \text{ ha}^{-1} \text{ year}^{-1}$ and the mean of turbidity observations was found to be 10.4 (Table 2). That is, the observed means from the trial were greater than past measurements for both MAI and turbidity. What inferences might be drawn from these results? Corporate stakeholders whose core interest is the company's profitability may be inclined to interpret the results as evidence supporting broader adoption of mechanical disturbance. Similarly, stakeholders concerned about environmental outcomes may see the results as a vindication of their view that the treatment will pose a serious threat to frog populations.

Intuitively, the likelihood that the sample means are consistent with reality is related to sampling intensity and the magnitude of variation in MAI and stream turbidity. A standard approach to evaluating sample size and variability is to use a statistical test involving null (H0) and alternative (H1) hypotheses. For MAI, the null and alternative hypotheses might be stated as,

- H0: The sample mean was drawn from a population with mean = $4.1 \text{ m}^3 \text{ ha}^{-1} \text{ year}^{-1}$ (i.e. the silvicultural treatment has no effect).
- H1: The sample mean was drawn from a population with mean $>4.1 \text{ m}^3 \text{ ha}^{-1} \text{ year}^{-1}$ (i.e. a one-tailed test for the hypothesis the silvicultural treatment improves growth); with corresponding hypotheses for turbidity.

Using a *t*-test, the test statistic is calculated using (Sokal and Rohlf, 1995):

$$t = \frac{\bar{x} - T}{s/\sqrt{n}}, \quad (1)$$

where \bar{x} is the sample mean, *T* the threshold reference against which the mean is compared, *s* the sample standard deviation and *n* is the sample size.

The *t*-test asks, what is the probability (*p*-value) of obtaining a value as large as the test statistic assuming the sample mean was drawn from a population having a mean of *T* (i.e. assuming the null hypothesis is correct). The convention employed in scientific inference is to reject the null hypothesis in favor of the alternative if the *p*-value is less than 0.05. The null is retained if

Table 2
Monitoring results from a hypothetical scenario examining the productive and environmental effects of a proposed change in silviculture

	Reference	Impact		
		Mean	S.D.	<i>n</i>
MAI	4.1	4.6	1.4	50
Turbidity	9.2	10.4	7.8	25

Productive effects are described by improvement in mean annual increment (MAI). Environmental effects refer to a postulated increase in turbidity leading to decline in a threatened frog species. Reference values are the means prior to silvicultural change. Impact values describe summary statistics from monitoring sites at which the silvicultural change has been trialed.

the p -value is greater than 0.05. The value of 0.05 is more or less arbitrary. It represents a burden of proof with which science is generally comfortable.

For MAI the test result is $t = 2.525$, $p = 0.007$. And for turbidity the result is $t = 0.769$, $p = 0.225$. These results might please the company's corporate stakeholders. The results show that the effect is a statistically significant improvement in MAI and a non-significant increase in turbidity.

Although common, the use of null hypothesis testing in this way is naïve and can lead to serious managerial mistakes. There are two key shortcomings. Firstly, the p -value in standard null hypothesis testing relates only to the probability of a Type I error. The result includes no indication of the probability of a Type II error (inferring no effect when in fact one exists). Secondly, statistical significance at $p < 0.05$ is often mistaken as providing information about the size or importance of the effect (Fidler, 2006). Unfortunately it does not. In fact, even trivially small effects can be statistically significant when very large sample sizes are involved.

The danger of ignoring these shortcomings can be illustrated by changing the sampling intensities in our scenario. Let us say that instead of sampling 50 inventory plots for MAI the company sampled only 15. And that instead of 25, the company made 120 observations of stream turbidity. (Let us also say the sample means and standard deviations remain unchanged from those shown in Table 2). In these circumstances, the inferences that arise from the t -test are reversed (Table 3).

The proposition that management decisions might rest on something as seemingly arbitrary as sampling intensity is alarming. But without consideration of effect size and Type I and Type II errors, naïve use of null hypothesis testing will tend to mislead. Such mistakes are commonplace, even in published scientific literature (Di Stefano et al., 2005; Fidler et al., 2006). In our scenario, it may be reasonable to assume that mechanical disturbance will enhance growth to some degree through improved aeration of the soil and an initial flush of enhanced nutrient availability. Similarly, it is reasonable to assume that mechanical disturbance will inevitably lead to some increase in stream turbidity, however small. Where statistical tests are conducted for any effect (no matter how miniscule) the tendency will be for larger sample sizes to report statistically

Table 3
The sensitivity of inferences to sample size

	n	p -Value	Inference
Case 1			
MAI	50	0.007	Reject H0: there is a significant increase in yield
Turbidity	25	0.225	Retain H0: there is no significant increase in turbidity
Case 2			
MAI	15	0.094	Retain H0: there is no significant increase in yield
Turbidity	120	0.047	Reject H0: there is a significant increase in turbidity

Two cases involving different sample sizes for two attributes of interest can give conflicting results. The apparent conflict arises because of the tendency to overlook statistical power when using null hypothesis significance testing.

significant results relative to smaller samples. In this sense, the use of the term 'significant' in statistical inference is unfortunate. Statistical significance is not equivalent to ecological, social or economic significance.

2.2. Strict application of statistical power

Acknowledgement of the shortcomings of standard approaches to null hypothesis significance tests encourages managers, stakeholders and auditors to think more critically about the use of data gathered in monitoring. In the context of claims that data support proposals that improve productive capacity, managers and Board members may reasonably ask, 'Has the interpretation been based on a commercially important effect size or on any effect size? What is the chance of inferring an important effect when in fact none exists (false optimism)? What is the chance of inferring no effect when in fact a commercially important effect exists (false pessimism)?' In an environmental and social context, community stakeholders will be especially interested in minimizing the likelihood of inferring no important impact when in fact one exists (false sense of security). Public relations managers will be concerned about the likelihood of false alarmism (inferring an important impact when in fact none exists). To formally and directly address these questions, power calculations are required.

What was the power of the two sampling scenarios explored above for MAI and turbidity? First we need to revisit the way we frame our hypotheses to take into account what might represent an important effect (as opposed to any effect). Let us say that, after accounting for costs in equipment and personnel to implement intensive mechanical disturbance, management identifies that MAI in the first 3 years needs to be at least $4.5 \text{ m}^3 \text{ ha}^{-1} \text{ year}^{-1}$ for the treatment to be cost-effective. For stream turbidity, an ecologist identifies 12.0 units as a threshold at which the rare frog species begins to be substantially adversely affected. The null and alternative hypotheses for MAI are now:

- H0: The sample mean was drawn from a population with mean = $4.1 \text{ m}^3 \text{ ha}^{-1} \text{ year}^{-1}$ (i.e. the silvicultural treatment has no effect).
- H1: The sample mean was drawn from a population with mean $\geq 4.5 \text{ m}^3 \text{ ha}^{-1} \text{ year}^{-1}$ (i.e. a one-tailed test for the hypothesis the silvicultural treatment improves growth enough to be cost-effective); again, with corresponding hypotheses for turbidity.

If we defer to scientific convention and accept a Type I error rate $\alpha = 0.05$, we can identify the value of the sample mean beyond which the null hypothesis will be rejected. Expressed algebraically:

$$\alpha = p\left(\bar{x} > \mu + z \frac{\sigma}{\sqrt{n}}\right), \quad (2)$$

where μ is the population mean associated with the null hypothesis, σ the population standard deviation (for which the sample standard deviation s is commonly used as an

estimate), and z comes from the standard normal distribution. For the one-tailed tests in our scenario $z = 1.645$ at $\alpha = 0.05$. Note that the value of z is equivalent to t for an infinite sample size, n , and that Eq. (2) is (a transposed) equivalent to Eq. (1).

For the case where MAI is estimated from $n = 50$, and the sample estimate of 1.4 for the population standard deviation is used, 4.43 is the critical value beyond which the null hypothesis is rejected. If H_0 is true, for a sample size of 50, the chance that the sample mean will exceed 4.43 is 5%.

A Type II error arises when the actual mean MAI is 4.5 (or greater), but the sample mean is less than 4.43. For the case where $n = 50$:

$$\beta = P\left(\bar{x} < 4.43 \mid \bar{x} \sim N\left(4.5, \frac{1.4}{\sqrt{50}}\right)\right) = 0.35.$$

That is, for a true population mean of 4.5, there is a 35% chance that the mean of a sample of 50 inventory plots will be less than 4.43, leading to the (incorrect) inference that the silvicultural treatment is ineffective in increasing yield. The power of the test is $1 - \beta = 0.65$, to be interpreted as a 65% chance of finding a difference of this size if it exists in the population. Table 4 provides results from equivalent calculations for the other t -tests involved in our two sampling scenarios, and Fig. 1 shows results graphically.

What is acceptable power? Community stakeholders would much prefer turbidity monitoring to have 99% power than 56% power. That is, they may regard as intolerable a 44% chance of inferring the new silvicultural practice is no threat to frogs

Table 4

Type II (β) error rates and power ($1 - \beta$) for two sampling intensities used to test productive and environmental effects of a proposed change in silviculture

	n	Critical value	β	$1 - \beta$	$\alpha:\beta$
Case 1					
MAI	50	4.43	0.35	0.65	0.14
Turbidity	25	11.77	0.44	0.56	0.11
Case 2					
MAI	15	4.69	0.70	0.30	0.07
Turbidity	120	10.37	0.01	0.99	4.51

The ratio of the Type I to Type II error rate ($\alpha:\beta$) refers to a fixed Type I error rate of 0.05. It indicates the relative likelihood of false rejection (α) or false retention (β) of the null hypothesis.

when in fact it is. A 1% chance of such an outcome would be far more palatable. Corporate stakeholders may regard sampling to be inadequate in both cases. A 35–70% chance of missing a commercially important improvement in yield may represent an intolerable opportunity cost to the company.

The power of any monitoring program or research project is proportional to the effect size to be detected (ES), the Type I error rate (α), the square root of the sample size (n), and the magnitude of variability of the population from which samples are taken (σ) (Fairweather, 1991). That is:

$$\text{power} \propto \frac{ES\alpha\sqrt{n}}{\sigma}.$$

In most circumstances, the most effective options available for increasing power are to increase sample size or accept a higher

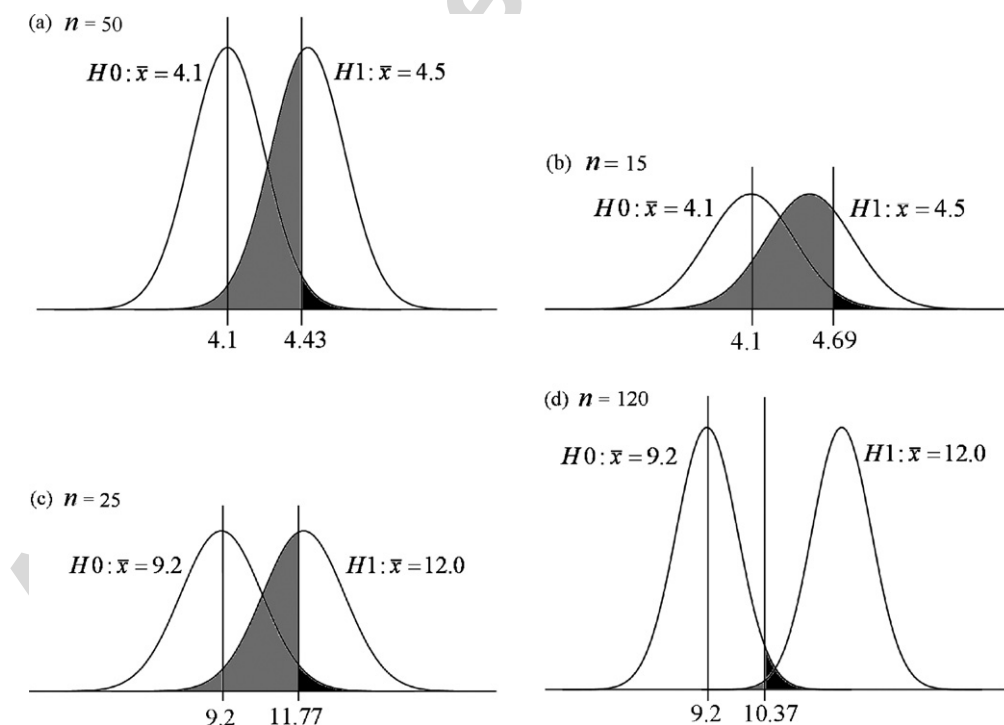


Fig. 1. Results of power analysis for contrasting sampling intensities measuring MAI and turbidity. The specified Type I error rate of $\alpha = 0.05$ is shown black in each case. Grey areas correspond to the probability of a Type II error, conditional on the alternative hypothesis being true. For MAI (a and b), curves show the distribution of means that would be obtained from null and alternative hypothesis populations with a standard deviation of 1.4. For turbidity (c and d) curves show distributions of means drawn from populations with a standard deviation of 7.8. Smaller sample sizes are characterized by flatter curves, leading to relatively higher chances of committing Type II errors.

Type I error rate. Improved measurement practices may also contribute through reducing variability associated with random measurement error. As with the specification of an effect size, the question of what is an acceptable error rate for Type I and Type II errors is not a decision that should be left to the arbitrary conventions of statisticians or technicians (Di Stefano, 2003). It involves judgments from managers, auditors and stakeholders on the extent to which they can tolerate false optimism, false pessimism, false alarmism or a false sense of security.

Mapstone (1995) and Di Stefano (2001) recommend using a ratio of α to β that reflects the relative costs of the two kinds of errors, and designing the monitoring program accordingly. For frog impacts associated with turbidity, Table 4 reports a ratio of false alarmism to false sense of security of 0.11 for the first case and 4.51 for the second. Those concerned with the public image of the company might be pleased with the former, whereas community stakeholders will be more comfortable with the latter. Levels of tolerance will vary among interested parties according to the values they seek to promote in forest management and according to the extent to which they incur the financial costs of collecting data.

Power calculations can be conducted before (a-priori) a study or monitoring program is undertaken or after (post hoc). A priori analyses commonly seek to estimate the sampling intensity required where the effect size to be detected has been identified, tolerable thresholds for Type I and Type II errors have been specified, and an estimate of the population's variance is available (Di Stefano, 2001; Foster, 2001). Post hoc analyses usually estimate the Type II error for whatever sample size has been employed. Employed correctly, they avoid using observed differences, instead relying on effect sizes that could have been specified a priori (Hoening and Heisey, 2001).

Formulae for power calculations vary according to the particular statistical test being used. In the example of a proposed change to silviculture explored in this section, we used a simple *t*-test to demonstrate the concepts. Burgman et al. (1998) and Walshe and Wintle (2006) give further forestry examples of simple power calculations. Complex problems require solutions involving computer simulation. The basis of these solutions involves randomization, whereby observed data are resampled or reshuffled many times to generate the sampling distribution (Quinn and Keough, 2002). While the method of constructing confidence intervals using randomization is different to the frequentist parametric approach, the core principles of statistical power still apply. Manly (1997) provides an overview of randomization methods and Crowley (1992) critiques applications in biology.

3. Confidence intervals as an alternative to power calculations

The formulation of power calculations is not intuitive. In the scenario above, the calculation requires us to specify a null hypothesis of no effect and an alternative hypothesis equivalent to an effect size deemed to be of commercial, social or environmental consequence. Values for α and β need to be specified. In a priori calculation, the number of samples needed

to satisfy these thresholds is calculated. In the context of auditing a standard, the objective of the calculations is to differentiate clearly compliance from non-compliance. In post hoc calculations, we seek to establish clearly whether our sampling design succeeded or failed to satisfy the burden of proof implicit in specified Type I and Type II error rates.

Clarity in auditing might be desirable, but for many people strict application of statistical power is not a natural or intuitive way to approach the problem. Part of the difficulty is the emphasis power calculations place on dichotomous outcomes. Compliance is made a black and white issue—a company either complies or fails to comply. This dichotomy may be unreasonable (and uninformative) where the subjective judgments necessary in defining effect sizes and tolerable Type I and Type II error rates are contested, arbitrary or vague.

In any case, communication may be problematic. Methodological rigor offers little insulation to stakeholder skepticism (Bocking, 2004). Challenges to the validity of audit outcomes are to be expected, even where power calculations are based on demanding error rates. In these circumstances, making power calculations communicative and accessible to stakeholders without technical expertise is likely to be difficult.

An alternative to strict interpretation and application of statistical power is the use of confidence intervals (CIs). CIs communicate the key elements of statistical power and *p*-values but are not bound to dichotomous outcomes (Gardner and Altman, 2000; Burgman, 2005; Fidler, 2006). The central feature of statistical power in the context of monitoring and auditing is that it remedies the predisposition of stakeholders to view equivocal data as supportive of their position. This feature is preserved with CIs.

In precise language, a 90% CI is the interval which will contain the true value on 90% of occasions if a study were repeated many times using samples from the same population. In practice, the interval is interpreted to be the plausible range for the true population value.

The graphical communication of confidence intervals emphasizes effect size and uncertainty associated with sampling and measurement error (Di Stefano, 2004). As is the case with power calculations, CIs communicate uncertainty associated only with sampling error. It cannot describe or control non-sampling error in design and conduct of monitoring programs. The width of confidence intervals depends on three factors—sample size (*n*), variability of the characteristic being monitored or studied (*s*), and the degree of confidence required (Gardner and Altman, 2000).

The degree of confidence required is akin to specifying error rates in power calculations, but not equivalent. Although 95% CIs are common in the scientific literature, there is no reason to apply the convention in management contexts. The confidence level used in any circumstance represents a trade-off between the costs of conservatism and the costs of over-confidence.

Returning to our example of proposed silvicultural change, lower and upper bounds for CIs are calculated using, $\bar{x} \pm [t_{1-\alpha/2} SE_{\bar{x}}]$, where $t_{1-\alpha/2}$ is the appropriate value from the *t* distribution with *n* – 1 degrees of freedom associated with a 'confidence' of 100(1 – α)%. The 90% CI for the mean

estimate of MAI = 4.6 in Case 1 is [4.27, 4.93]. The terms used to calculate CIs are the same as those used to calculate the test statistic in a *t*-test. The element of power calculations missing from CIs is effect size, but this can be added graphically.

Results for MAI and turbidity under Case 1 and Case 2 sampling are presented in Fig. 2. The intervals in Fig. 2 provide the following interpretations: (a) The silvicultural change improves MAI, but we are unsure if the magnitude of the improvement is commercially viable. (b) Monitoring is largely inconclusive. The silvicultural change could result in a decrease in turbidity or an increase large enough to negatively impact frog populations. (c) Monitoring is largely inconclusive. The silvicultural change could result in a decrease in MAI or an increase large enough to be commercially attractive. (d) The silvicultural change increases turbidity, but the magnitude of the increase is too small to substantially affect frog populations. The qualitative insights are consistent with power calculations, but results are presented in a way that promotes intuitive understanding.

The outcomes associated with our example of a proposed change in silviculture (Fig. 2) are not exhaustive. Where a clear threshold for effect size can be specified, outcomes include clear evidence of an important effect, clear evidence of no important effect, and a non-conclusive interval. Limitations in our technical understanding of what might constitute an important effect size mean that defining clear thresholds can be difficult. Value judgments will also contribute legitimately to perspectives on what is considered an important effect. Where the perspectives of scientists, managers and stakeholders cannot be reconciled in specifying a threshold for assessing compliance, a lower and upper bound can be accommodated.

In Fig. 3, the dashed line represents zero effect and the grey box is delimited by the lower and upper bound of the effect size we are interested in detecting. For the sake of illustration, we restrict our exploration to circumstances where an exceedence of the bounded effect size represents an undesirable or intolerable outcome. Note that inclusion of ‘zero effect’ is often not necessary. We include it here to emphasize the

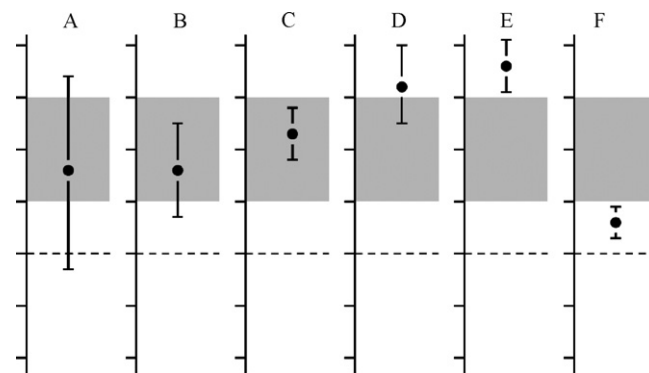


Fig. 3. Six possible outcomes of plotting confidence intervals against a vaguely defined threshold. The grey box is delimited by lower and upper bounds of the effect size we are interested in assessing. The dashed horizontal line refers to no effect. See text for details.

importance of discriminating between an important effect and any effect. There are six possible outcomes that may imply different management responses in the context of continuous improvement under a forestry standard: (A) is largely uninformative. There is a clear need for greater precision through increasing the sampling intensity of monitoring. In the interim, managers may be required to implement major risk mitigation measures to avoid the possibility of intolerable outcomes. (B) suggests the effect is intolerable for stakeholders associated with the lower bound of the effect size, but it is possible the true effect is less than this threshold. Managers are likely to be motivated to seek improved clarity through increased sampling intensity. In the meantime, minor risk mitigation measures may be needed. (C) is a relatively precise description of effect size, falling wholly within stakeholder bounds for tolerable effect size. There is no need to increase sampling intensity. Minor risk mitigation measures may be warranted to address the concerns of a subset of stakeholders. (D) suggests a large effect that is likely to be intolerable to all stakeholders, but may fall below the upper bound. Major risk mitigation may be required. Managers could increase sampling

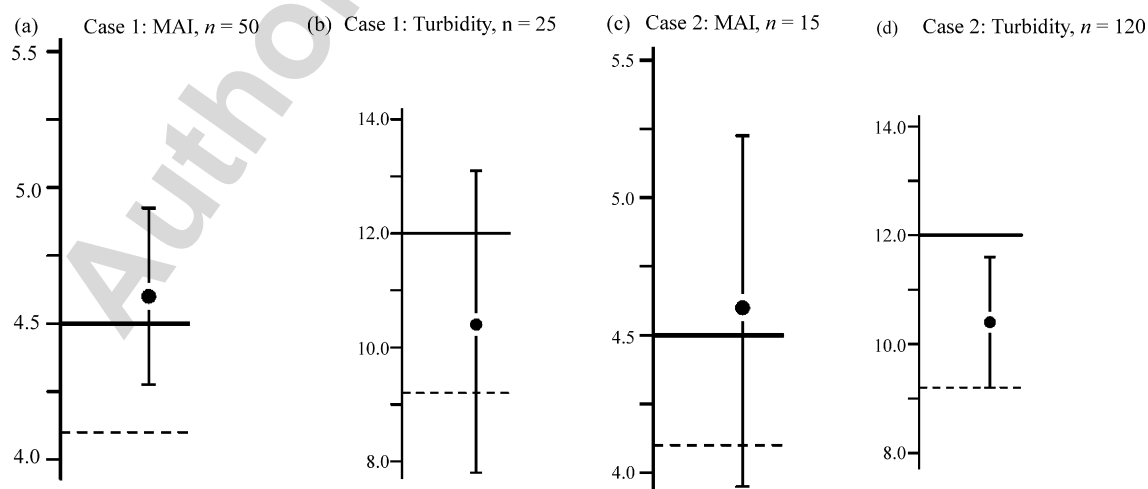


Fig. 2. 90% confidence intervals for changes in MAI and stream turbidity associated with a change in silviculture. The dashed horizontal line refers to the background MAI or turbidity (i.e. no effect) and the bold continuous horizontal line is the effect size we are interested in detecting.

intensity to clarify whether lesser mitigation measures may be more appropriate. (E) shows clear evidence of an important effect requiring major mitigation measures. (F) indicates clear evidence of an unimportant effect. The intensity of monitoring may be reduced.

4. Discussion

We recommended confidence intervals as graphical means of facilitating evidence-based continuous improvement, providing the basis for informed discussion on what might (a) constitute compliance and non-compliance and (b) the extent to which a monitoring program is able to distinguish compliance from non-compliance.

Monitoring programs that clearly differentiate circumstances in which management complies or does not comply typically demand intensive sampling (Mapstone, 1995). Our guess is that with typical budgets dedicated to monitoring, there will be many instances where the upper and lower bounds of CIs encompass thresholds. That is, intervals will commonly fail to distinguish compliance from non-compliance. This is not a weakness of the approach we advocate. Rather, the inadequacy of resources dedicated to monitoring is made plain to managers, auditors and stakeholders. Credibility of the notions of evidence-based continuous improvement and decision-making would be substantially improved if a consequence of broader awareness of statistical power and the use of CIs in reporting was more efficient allocation of resources to monitoring.

However, our expectation is that it will be cost-prohibitive for managers to allocate sufficient resources to monitoring all criteria under a standard in a way that satisfies demanding Type I and Type II error rates. We suggest a tiered approach to demonstrating compliance, whereby criteria considered most important by managers and stakeholders are given a stricter burden of proof than criteria of lesser importance. For example, if stakeholders are more concerned about water pollution than visual amenity, a conservative interval (say 95%) and/or more stringent threshold might be required for indicators associated with water quality. Greater tolerance for visual impacts may suggest more lenient thresholds and less conservative intervals (say 50%).

The essential role of monitoring in the context of audited standards is to test the validity of assertions regarding impact that are implicitly embedded in the policies and procedures of management. Conceptually, the requirement to undertake rigorous and intensive monitoring is inversely proportional to the extent to which a precautionary approach has been adopted. A risk-averse or precautionary approach implies low likelihood of a negative impact, and investment in monitoring may be a lesser imperative.

The unique circumstances of any single organization accredited under a standard need to be recognized. We suggest subjective risk assessment be used as an efficient and effective means of discerning high and low priority criteria within a standard. Risk assessment has been a common element of planning and management in medicine, engineering and process industries for several decades, and prescriptive

protocols have been developed for its application (e.g. ICE/FIA, 1998; ISO 14971-1, 1998; AS/NZS, 2004).

Essentially, risk assessment considers what might compromise identified values. Few would argue that this is a sensible and reasonable thing to do in planning and management. But when done poorly, risk assessment can be seen as a process which attempts to make the illegitimate exercise of power by vested interests and scientists acceptable (O'Brien, 2000). It has the potential to alienate stakeholders by placing societal perspectives in the hands of the technical people who conduct the assessments (Fischer, 2000). Stakeholders should be made central participants of risk assessments.

The Australian Standard for Risk Management (AS/NZ 4360; AS/NZS, 2004) details protocols for identifying values and hazards and for estimating the magnitude of risk associated with each hazard. Inferring the relative risk of hazards involves use of a matrix of consequence and likelihood. In consultation with stakeholders, outcomes depend on the capacity of the analyst to (a) identify relevant values, (b) elicit an exhaustive list of potential hazards, and (c) use subjective judgment for each potential hazard to estimate the likelihood that an event will occur and the severity of its consequences. Priorities for monitoring (and management intervention) are those that are considered higher risk and those for which there is considerable variation in perspectives among stakeholders.

A hypothetical example of outcomes from a risk assessment is shown in Table 5. The response of auditors and managers to this information may involve stipulation of a conservative confidence level and concomitant preferential allocation of monitoring resources to the extent of regeneration failure, the incidence of the fungal pathogen, and the concentration of contaminants in streams. Relatively low intensity monitoring may be undertaken for soil nutrients, insect pests, perceptions of visual amenity, and the magnitude of any decline in hollow-bearing trees. Monitoring and review of procedures for low risk hazards (protecting remnant vegetation and the extent and impact of spray drift) may be considered unnecessary.

Although this simple approach to risk assessment has a number of frailties (Burgman, 2005) it is important to recognize its advantages. It is simple and fast, accounts for probability of harm and magnitude of harm, communicates environmental and social risk in the same language used for financial risk, provides an informal means of combining data and expert

Table 5
Hypothetical example of outcomes of a risk assessment conducted using the Australian Standard (AS/NZ 4360; AS/NZS, 2004)

Hazard	Risk
Loss of soil nutrients	Medium
Reduction in hollow-bearing trees	Low-medium
Spray drift from aerial application of pesticides	Low
Loss of visual amenity	Low-medium
Intense insect attack	Medium
Remnant vegetation protection	Low
Regeneration failure	High
Spread of fungal pathogen	Low-high
Stream pollution	Low-high

judgment, and it provides an auditable record of priorities (Hart et al., 2005).

Our view is that the prospects for evidence-based continuous improvement will be markedly enhanced under systems that include the complementary insights afforded by risk assessment, statistical power and the use of confidence intervals in communicating performance.

Acknowledgments

This work was funded by the Forest & Wood Products Research & Development Corporation (FWPRDC) through a grant made available to the Australian Forestry Standard Ltd. The FWPRDC is jointly funded by the Australian forest and wood products industry and the Australian Government. For comments received on a draft we thank Julian Di Stefano, Mark Edwards, David Flinn, Hans Drielsma, Ross Peacock, Wayne Hammond, John Wiedemann, Erwin Epp, Marks Nester, Kevin Swanepoel and three anonymous reviewers.

References

- ANZECC/ARCMANZ, 2000. Australian Guidelines for Water Quality Monitoring and Reporting. Australian and New Zealand Environment and Conservation Council/Agriculture and Resource Management Council of Australia and New Zealand. Australian Government Publishing Service, Canberra.
- AS/NZS, 2004. Risk Management (AS/NZS 4360:2004). Standards Australia International, Sydney.
- Bocking, S., 2004. Nature's Experts. Science, Politics and the Environment. Rutgers University Press, New Brunswick.
- Burgman, M.A., 2005. Risks and Decisions for Conservation and Environmental Management. Cambridge University Press, Cambridge.
- Burgman, M.A., Ades, P., Hickey, J., Williams, M., Davies, C., Maillardet, R., 1998. Methodological guidelines for the utility and statistical validity of survey and monitoring programs. Report No. PN98.803. Forest & Wood Products Research & Development Corporation, Victoria.
- Crowley, P.H., 1992. Resampling methods for computer-intensive data analysis in ecology and evolution. *Annu. Rev. Ecol. Syst.* 23, 405–447.
- Di Stefano, J., 2001. Power analysis and sustainable forest management. *For. Ecol. Manage.* 154, 141–153.
- Di Stefano, J., 2003. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Funct. Ecol.* 17, 707–709.
- Di Stefano, J., 2004. A confidence interval approach to data analysis. *For. Ecol. Manage.* 187, 173–183.
- Di Stefano, J., Fidler, F., Cumming, G., 2005. Effect size estimates and confidence intervals: an alternative focus for the presentation and interpretation of ecological data. In: Burk, A.R. (Ed.), *New Trends in Ecology Research*. Nova Science Publications, New York, pp. 71–102.
- Fairweather, P.G., 1991. Statistical power and design requirements for environmental monitoring. *Aust. J. Mar. Freshwater Res.* 42, 555–567.
- Fidler, F., 2006. From statistical significance to effect estimation. Statistical reform in psychology, medicine and ecology. PhD Thesis. Department of History and Philosophy of Science, University of Melbourne.
- Fidler, F., Burgman, M.A., Cumming, G., Buttrose, R., Thomason, N., 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv. Biol.* 20, 1539–1544.
- Field, S.A., Tyre, A.J., Jonzén, N., Rhodes, J.R., Possingham, H.P., 2004. Minimizing the costs of environmental management decisions by optimizing statistical thresholds. *Ecol. Lett.* 7, 669–675.
- Fischer, F., 2000. Citizens, Experts, and the Environment. Duke University Press, Durham.
- Foster, J.R., 2001. Statistical power in forest monitoring. *For. Ecol. Manage.* 151, 211–222.
- Fowler, J., Cohen, L., Jarvis, P., 1998. *Practical Statistics for Field Biology*, 2nd ed. John Wiley & Sons, Chichester.
- FSC, 2000. FSC Principles and Criteria for Forest Stewardship (FSC-STD-01-001). Forest Stewardship Council, Bonn.
- Gardner, M.J., Altman, D.G., 2000. Confidence intervals rather than *P* values. In: Altman, D.G., Machin, D., Bryant, T.N., Gardner, M.J. (Eds.), *Statistics with Confidence. Confidence Intervals and Statistical Guidelines*. 2nd ed. BMJ Books, London.
- Hart, B., Burgman, M., Webb, A., Allison, G., Chapman, M., Duivenvoorden, L., Feehan, P., Grace, M., Lund, M., Pollino, C., Carey, J., McCrea, A., 2005. Ecological Risk Management Framework for the Irrigation Industry. Report to the National Program for Sustainable Irrigation (NPSI) Water Studies Centre, Monash University, Clayton, Australia.
- Hoenig, J.M., Heisey, D., 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Stat.* 55, 19–24.
- ICE/FIA, 1998. RAMP: Risk Analysis and Management for Projects. Institution of Civil Engineers and the Faculty and Institute of Actuaries. Thomas Telford, London.
- ISO 14971-1, 1998. International Standard 14971-1. Medical Devices—Risk Management. Part 1. Application of Risk Analysis. International Organisation for Standardization, Geneva.
- Ludwig, D., Hilborn, R., Walters, C., 1993. Uncertainty, resource exploitation, and conservation: lessons from history. *Science* 260, 17–36.
- Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd ed. Chapman and Hall, London.
- Mapstone, B.D., 1995. Scaleable decision rules for environmental impact studies: effect size, Type I and Type II errors. *Ecol. Appl.* 5, 401–410.
- McCarthy, M.A., 2007. Bayesian Methods for Ecologists. Cambridge University Press, Cambridge.
- Montreal Process, 1999. Criteria and Indicators for the Conservation and Sustainable Management of Temperate and Boreal Forests. 2nd ed. ISBN: 0-662-29009-7, 16 pp.
- Morgan, M.G., Henrion, M., 1990. Uncertainty. A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, Cambridge.
- O'Brien, M., 2000. Making Better Environmental Decisions: An Alternative to Risk Assessment. MIT Press, Cambridge, Massachusetts.
- Philip, M.S., 1994. Measuring Trees and Forests, 2nd ed. CABI Publishing, Wallingford.
- Prabhu, R., Ruitenbeek, H.J., Boyle, T.J.B., Colfer, C.J.P., 2001. Between voodoo science and adaptive management: the role and research needs for indicators of sustainable forest management. In: Raison, R.J., Brown, A.G., Flinn, D.W. (Eds.), *Criteria and Indicators for Sustainable Forest Management*. CABI Publishing, Wallingford, pp. 39–66.
- Quinn, G.P., Keough, M.J., 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.
- Raison, R.J., Flinn, D.W., Brown, A.G., 2001. Application of criteria and indicators to support sustainable forest management: some key issues. In: Raison, R.J., Brown, A.G., Flinn, D.W. (Eds.), *Criteria and Indicators for Sustainable Forest Management*. CABI Publishing, Wallingford, pp. 5–18.
- Sokal, R.R., Rohlf, F.J., 1995. *Biometry*, 3rd ed. Freeman, San Francisco.
- Standards Australia, 2003. Interim Australian Standard. The Australian Forestry Standard (AS 4708 (Int)—2003) Australian Forestry Standard Steering Committee, Canberra.
- Suter, G.W., 1993. *Ecological Risk Assessment*. Lewis, Boca Raton.
- Tversky, A., Kahneman, D., 1971. Belief in the law of small numbers. *Psychol. Bull.* 76, 105–110.
- Walshe, T., Wintle, B., 2006. Guidelines for communicating performance against standards in forest management. Forest and Wood Products Research and Development Corporation Report PN06.4012. <http://www.fwprdc.org.au/content/pdfs/new%20pdfs/PN06.4012.pdf>.