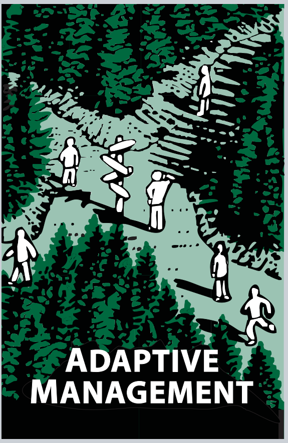
LAND MANAGEMENT HANDBOOK

42

Statistical Methods for Adaptive Management Studies



1998

BRITISH
COLUMBIA

Ministry of Forests
Research Program

Statistical Methods for Adaptive Management Studies

Vera Sit and Brenda Taylor (Editors)

BRITISH
COLUMBIA

Ministry of Forests
Research Program

Citation

Sit, V. and B. Taylor (editors). 1998. Statistical Methods for Adaptive Management Studies. Res. Br., B.C. Min. For., Res. Br., Victoria, BC, Land Manage. Handb. No. 42.

Canadian Cataloguing in Publication Data
Main entry under title:
Statistical methods for adaptive management studies

(Land management handbook ; 42)

Includes bibliographical references: p.
ISBN 0–7726–3512–9

1. Forest management – Statistical methods.
2. Forest management – British Columbia. I. Sit, Vera.
II. Taylor, Brenda, 1962– . III. British Columbia. Ministry
of Forests. Research Branch. IV. Series.

SD387.S73S72 1998 634.9’2’072 C98–960081–5

Prepared for:
B.C. Ministry of Forests
Research Branch
PO Box 9519, Stn Prov Govt
Victoria, BC V8W 9C2

Published by
B.C. Ministry of Forests
Forestry Division Services Branch
Production Resources
595 Pandora Avenue
Victoria, BC V8W 3E7

Copies of this and other Ministry of Forests
titles are available from:
Crown Publications Inc.
521 Fort Street
Victoria, BC V8W 1E7

Send comments to: Vera Sit, B.C. Ministry of Forests, Research Branch,
PO Box 9519, Stn Prov Govt, Victoria, BC V8W 9C2

Ministry of Forests Publications Internet Catalogue: www.for.gov.bc.ca/hfd

ACKNOWLEDGEMENTS

This report is the combined effort of a number of people. We would like to thank the chapter authors not only for their written contributions, but also for discussions that helped to identify and refine the contents of the report. Many thanks to our reviewers: Balvinder Biring, Ian Cameron, Phil Comeau, Nola Daintith, Darrell Errico, Roger Green, George Harper, Russ Horton, Val LeMay, Gordon Nigh, Peter Ott, Rick Page, Martin Raphael, Michael Pitt, Pasi Puttonen, Doug Steventon, Michael Stoehr, Jeff Stone, Keith Thomas, Chris Thompson, Ian Thompson, and Rita Winkler. Comments and suggestions from the reviewers strengthened each chapter and the report as a whole. We would also like to extend our appreciation to Brian Nyberg for his full support and continued encouragement throughout the project, and to Paul Nystedt for assistance with the report's publication. Funding for this report was provided by the B.C. Ministry of Forests, Forest Practices Branch.

Vera Sit
Brenda Taylor

Victoria, B.C.
March 1998

Judith L. Anderson
Simon Fraser University, School of Resource and Environmental Management,
Burnaby, BC V5A 1S6

Wendy A. Bergerud
B.C. Ministry of Forests, Research Branch,
P.O. Box 9519, Station Provincial Government, Victoria, BC V8W 9C2

Bruce G. Marcot
U.S. Forest Service, Department of Agriculture, Pacific Northwest Research Station,
1221 SW Yamhill Street, Suite 200, Portland, OR 97208-3890

Amanda F. Linnell Nemec
International Statistics and Research Corporation,
P.O. Box 496, Brentwood Bay, BC V8M 1R3

J. Brian Nyberg
B.C. Ministry of Forests, Forest Practices Branch,
P.O. Box 9513, Station Provincial Government, Victoria, BC V8W 9C2

Randall M. Peterman
Simon Fraser University, School of Resource and Environmental Management,
Burnaby, BC V5A 1S6

Calvin N. Peters
ESSA Technologies Ltd.,
Suite 300, 1765 West 8th Avenue, Vancouver, BC V6J 5C6

William J. Reed
University of Victoria, Department of Mathematics and Statistics,
P.O. Box 3045, Victoria, BC V8W 3P4

Richard D. Routledge
Simon Fraser University, Department of Mathematics and Statistics,
Burnaby, BC V5A 1S6

Carl J. Schwarz
Simon Fraser University, Department of Mathematics and Statistics,
Burnaby, BC V5A 1S6

Vera Sit
B.C. Ministry of Forests, Research Branch,
P.O. Box 9519, Station Provincial Government, Victoria, BC V8W 9C2

G. John Smith
Geo. John Smith Statistical Consulting Services,
2781 Point Grey Road, Vancouver, BC V6K 1A4

Brenda Taylor
B.C. Ministry of Forests, Forest Practices Branch,
P.O. Box 9513, Station Provincial Government, Victoria, BC V8W 9C2

TABLES

FIGURES

1 STATISTICS AND THE PRACTICE OF ADAPTIVE MANAGEMENT

J. BRIAN NYBERG

Abstract

As adaptive management becomes more widely recognized as a foundation element of good land stewardship, many resource professionals are attempting to extend its theories and principles into common practice. They wish to conduct powerful management experiments, to monitor the outcomes effectively and efficiently, and to use the resulting data to make reliable inferences for future decisions. Most managers, however, have little formal training in the application of experimental design and statistics to the problems that they want to address through adaptive management. This chapter sets the stage for the in-depth discussions of key aspects of statistics in adaptive management that are presented in subsequent chapters. It includes a working definition of adaptive management, demonstrates the value of the application of adaptive management to forestry issues, and explains some of the differences between research studies and adaptive management techniques.

1.1 Introduction

The concept of adaptive management (Holling [editor]1978) is steadily gaining wider acceptance in forestry, especially in Canada and the United States (e.g., Schmiegelow and Hannon 1993; Bormann et al. 1994; Nyberg and Taylor 1995; Covington and Wagner [technical coordinators] 1996; MacDonald et al. 1997). As a hybrid of scientific research and resource management, adaptive management blends methods of investigation and discovery with deliberate manipulations of managed systems. Through observation and evaluation of the ways that human interventions affect managed systems, new knowledge is gleaned about system interactions and productive capacities. This new knowledge is then applied to future decisions in a cycle of continuous improvement of policies and field practices.

Adaptive management has somewhat different goals from research and presents challenges that differ both in scope and nature from those posed by typical forest research studies. Consequently, designing and analyzing adaptive management studies involves more than simply transferring research techniques to management problems. Scientists can play an important role in adaptive management (Walters 1986), but it is local resource professionals who must become the "adaptive managers" if the promise of the concept is to be realized through its application to a large proportion of forested lands. As part of their everyday jobs, these managers (see above for clarification of the term) must be able to implement or even design studies that produce reliable information about issues that concern or challenge them.

This suggests that resource managers might need to use statistics in such studies. Few field-level managers, however, have experience in applying experimental designs and statistical methods, even in

The terms "manager" and "researcher" are used here in the following senses:

Managers (or resource managers) are responsible for making or approving decisions about forest resource use and conservation. Although there are exceptions, resource managers in British Columbia usually have university or technical institute training to the level of the undergraduate degree or diploma, and are often registered as professional foresters, agrologists, engineers, geoscientists, or biologists. Resource managers are usually employed by government agencies or private forest companies. To understand the main ideas in this report, and to be effective in implementing adaptive management, managers should have a basic academic background in statistics and subsequent field experience in making or contributing to complex resource decisions.

Researchers are applied scientists, usually from government agencies or universities, who are responsible for conducting scientific studies of forest ecology and management. Their goals include both furthering knowledge of forests and explaining how human actions affect forests. In addition to their expertise in forestry or related disciplines, researchers usually have post-graduate training in statistical methods and experimental design. To benefit fully from this report, however, they should also have considerable experience in conducting forest research.

situations suited to the classical statistical techniques taught in most universities and colleges. Furthermore, the characteristics of some adaptive management studies make them unsuitable for many familiar designs and methods, including analysis of variance (ANOVA). Alternative approaches such as Bayesian statistics and meta-analysis can be helpful in some of these problematic cases, but most resource managers are not familiar with these approaches.

To be informative and efficient, adaptive management projects must be led by people who know what options for study designs and analyses are available, and the relative strengths and weaknesses of each. This is a reasonable if ambitious objective for resource managers, whose role in adaptive management usually includes articulating questions, selecting among alternative courses of action, and then implementing those actions. For the researchers and biometricians who often advise managers on the details of study designs, sampling, and analysis, a more comprehensive understanding of the various statistical techniques is required.

This report has been designed as a guide to statistical methods appropriate for adaptive management studies, with material that should interest both managers and researcher scientists. It should serve as an introduction for some resource managers and a refresher for others on statistical methods, their strengths and weaknesses, and their suitability for studies of different types of management problems. For researchers and biometricians, it should provide a refresher on classical (familiar) methods, an introduction to less familiar methods, and a discussion of the typical challenges that will be faced in applying both to the unfamiliar situations of adaptive management. Although all the methods discussed here have been previously described in other texts and reports, that material is widely scattered in the literature and is thus not easily available to forestry practitioners. This report brings them together under one cover and deals directly with their application to adaptive management of forests.

The design of studies and analysis of data—the themes of this report—are only two components of the much larger topic of adaptive management. The following section explains the procedural framework of adaptive management. For information on other aspects, including conceptual foundations and implementation, refer to Holling (editor, 1978), Walters (1986), Lee (1993), Gunderson et al. (1995), and Taylor et al. (1997). In addition, Taylor et al. (1997) include a comprehensive list of other references.

1.2 Towards a Working Definition

Adaptive management is a systematic process for continually improving management policies and practices by learning from the outcomes of operational programs. Its most effective form—"active" adaptive management—employs management programs that are designed to experimentally compare selected policies or practices, by evaluating alternative hypotheses about the system being managed. The key characteristics of adaptive management include:

- acknowledgement of uncertainty about what policy or practice is "best" for the particular management issue;
- thoughtful selection of the policies or practices to be applied;
- careful implementation of a plan of action designed to reveal the critical knowledge;
- monitoring of key response indicators;
- analysis of the outcome in consideration of the original objectives; and
- incorporation of the results into future decisions.

Increasing use of the term "adaptive management" by different agencies in different settings (e.g., Lancia et al. 1996; Namkoong 1997) has spawned various interpretations and misinterpretations of its meaning. Consequently it is for many little more than a fuzzy concept. To bring the concept into sharper focus and to encourage a shared understanding of adaptive management among resource professionals in British Columbia, Nyberg and Taylor (1995) proposed the definition listed in the text above.

This definition suggests that adaptive management must comprise an organized sequence of activities. The sequence begins with a thorough analysis of the problem being faced and then proceeds to the creation of a management plan that is designed to speed learning about the system. It is not complete until the planned management actions have been implemented, measured, and evaluated; and the resulting new knowledge has been fed back into the decision-making process to aid in future planning and management. This sequence of steps can be summarized as a six-step process: (1) problem assessment, (2) project design, (3) implementation, (4) monitoring, (5) evaluation, and (6) adjustment of future decisions.

The sequence may need to be repeated in a continuing learning cycle if uncertainties remain unresolved or new ones appear.

This report deals mainly with the second, fourth, and fifth steps in the adaptive management process, namely the design (thoughtful selection) of practices to be studied, the measurement (monitoring) of responses, and the evaluation (analysis) of results.

1.3 Experiments in Adaptive Management

Adaptive management can take two different modes: active and passive (Walters and Holling 1990). A critical feature of both modes is thorough exploration, often through a modelling or "gaming" process, of the potential effects of policies or practices that are being considered for implementation. In passive applications only one policy or practice is explored, whereas in active adaptive management multiple options are compared and contrasted. In both cases subsequent management activities reveal, through monitoring and evaluation of their results, the accuracy or completeness of the earlier predictions. These deliberately designed activities are "experiments" in the broad sense of the term; that is, deliberate tests or trials intended to provide information about the response of the system of interest.

The notion of experimentation is central to adaptive management. As Lee (1993, p. 9) puts it, "Adaptive management...embodies a simple imperative: policies are experiments; *learn from them.*" In fact, experimentation is the element that ultimately distinguishes adaptive management and experimental research from other approaches to learning about nature. These other approaches, including the retrospective and observational studies described in later chapters of this report, can contribute helpful knowledge to later adaptive management work, but they are not themselves adaptive management because they do not include deliberately planned experimental manipulations.

Experimentation is considered at some length in this report, but it is defined here quite broadly compared to many scientists' concept of a scientific experiment. For reasons of scale, expense, and others, adaptive management experiments will not always include controls, replication, multiple treatments, randomization, or other features commonly expected of traditional scientific research. Nevertheless, those designing adaptive management experiments should strive to balance practicality with rigour so as to provide reliable information in a timely and cost-efficient manner. As part of the design process it is also critical to consider the statistical methods that will be used to analyze the resulting data. The following chapters describe methods that can be used to enhance the value of data from studies that pose some of the design problems listed above.

1.4 Need for Adaptive Management

Uncertainty drives adaptive management (Walters 1986). There would be little need to develop new policies or methods if managers were dealing with stable, predictable ecological and social systems. The outcomes of management programs could be reliably predicted, and standard practices could be taught to each generation of young professionals. Adaptive management and other approaches for dealing with uncertainty would be of little value.

Resource managers, however, do not live in such a world (Hilborn 1987). Uncertainties are pervasive in their work. The major categories of uncertainty that trouble managers when they consider the future are:

- natural environmental variability (e.g., weather, fire, earthquakes, avalanches, volcanoes, stream flows, genetic composition of species, animal movements);
- human impacts on the environment through global climate change, new technology, and the growing population;
- lack of knowledge about most aspects of the ecosystems being managed; and
- variations in social and political goals expressed as varying budgets, shifting policy directions, and changing demands for commodities, services, and aesthetic values from forests.

Given that resource managers and policy makers are faced with such difficult challenges, what can they do? Scientific research is one avenue for addressing the problem of lack of knowledge, but research programs often take years to organize, carry out, and report results. Meanwhile, resource management decisions continue to be made and forests continue to fall and regenerate under human hands. Money and expertise for research in forestry and other natural resource disciplines continue to be constrained at levels far below those needed to address many important issues. Furthermore, scientific research is limited in the types of questions it can answer because many forestry practices have cumulative effects that are only apparent at scales of time, space, or both that are not amenable to investigation through traditional experimental research. For example, it is impossible to use classical experimental methods employing controls and replicated treatments to determine the effects of forestry on wildlife that use huge areas, such as caribou, or that are threatened with extinction or local extirpation, such as spotted owls.

When research, education, or personal experience fail to provide information needed for difficult decisions, managers typically turn to professional opinion followed by unstructured trial-and-error management. This approach to learning is often inefficient and unreliable. Unless management alternatives are carefully thought out and attention is paid to potentially confounding factors such as biases, random errors, and unmeasured influences of weather, site, or other factors, it is often impossible to say what really caused any observed response. This can lead to "myths" being accepted widely due to the strongly held opinions of one or a few people —opinions that are later found to be wrong.

For example, poorly conceived and unsuccessful field trials may have been the genesis of the formerly strong bias among foresters in central British Columbia against partial cutting in high-elevation stands of spruce (*Picea* spp.) and subalpine fir (*Abies lasiocarpa*). Until a few years ago, many believed that partial cutting was unsuitable in any and all spruce-fir stands. This belief was based largely on reports that stands that had been partially cut before 1970 had all been subsequently windthrown or infested by insects or disease. Recent partial cutting trials have shown, however, that spruce-fir forests can be windfirm and healthy (for several years, so far) if the harvest intensity and site and stand conditions are appropriate.

In contrast to the basic trial-and-error approach, adaptive management is a much more organized and powerful approach to learning from experience. Its greatest contribution to learning may lie in the notion of making explicit predictions of the expected outcomes of management actions, then comparing actual outcomes to the predictions before adjusting subsequent actions and the models used to make the initial predictions. By designing management actions as experiments stronger inferences can be drawn from their outcomes, reducing the chance of generating false notions about forest functions and impacts. Other potential benefits of adaptive management include more reliable answers to questions about effects of forestry over large geographic areas and long time frames; insight into the causes and processes of system responses; shared understanding among scientists, policy makers, and managers; systematic resolution of conflicts over policies and practices; and efficient use of staff and budgets to address clear objectives (Holling [editor] 1978; Lancia et al. 1996; Taylor et al. 1997). All of these benefits contribute to accelerated learning and to the ultimate goal of improved decisions and forest management in future.

As an example of a potential application of adaptive management, consider the problem that resource managers face when they examine the question of how to protect water quality and downstream fish habitat in small headwater streams, while still allowing some timber harvesting to take place nearby. This situation creates a common and difficult problem in areas of British Columbia where small streams are numerous, slopes are steep, and timber values are high. The weight of expert opinion and of evidence from larger streams suggests that some streamside vegetation must be retained to provide shade and leaf litter, prevent sedimentation, and prevent degradation of bank and channel structure. If large trees are left standing in a narrow (<30 m) riparian strip after logging of areas near small streams, however, they are often blown down by wind, which may cause serious soil disturbance and bank damage as their root wads pull up. The loss of potential timber revenue in the fallen trees is exacerbated by the risk of insects (especially bark beetles) colonizing riparian blowdown and infesting other stands. Therefore, simply leaving narrow reserve zones of unlogged timber along all headwater streams is often not economically or ecologically appropriate.

Because past experience and scientific research have not resulted in a reliable approach to managing riparian areas adjacent to the smallest streams, what can resource managers and scientists do? If the situation allowed, the easy answer might be to cease all logging within, say, 50 m of such streams. But agency direction and societal demands do not allow such a "hands off" approach. Resource managers could instead postpone logging in all such areas while waiting for more intensive research to be done; or they could simply pick one or more methods and carry on logging while hoping for the best. Neither of these alternatives is a suitable response in times of restricted research funding and rising public expectations for resource stewardship.

By designing management actions as experiments, as part of adaptive management, logging could proceed and at the same time yield important information on key indicators of stream conditions, fish and wildlife habitat, and responses of trees and other vegetation. Through modified logging operations managers could apply several alternative treatments, such as various widths of unharvested reserve zones or different degrees of partial cutting of riparian zones, or both. Among other benefits, such a program might reveal the cumulative effects on watershed dynamics and wildlife habitat of large and widespread treatments—effects that could not be studied in a more traditional, small-scale research experiment.

Because adaptive management has attributes common to scientific research and sometimes draws on research techniques such as controlled experiments, some may assume that an ambitious program of adaptive management would obviate the need for research. They would be wrong, for intensive research is far more suited than adaptive management to answering some questions. Intensive research can produce deeper knowledge of selected system processes, such as mechanisms of physiological response in seedlings exposed to varying temperature and moisture regimes, than could adaptive management. It may also be the only approach suitable for sorting out the interacting effects of a number of factors on some dependent variable. This in-depth knowledge may be crucial for building models used to forecast how the overall system will respond to management.

Adaptive management is most suited to selecting amongst alternative courses of action, such as different partial cutting treatments that could be applied to a particular site and stand type. It can also be helpful for testing the modelled responses of managed systems against real-world results of management, across a much wider range of conditions than could any practical program of intensive research.

In fact, research and adaptive management complement each other, so that the application of both approaches to a problem will almost certainly lead to better results than use of either alone. Adaptive management can reveal management "surprises"; research can help to explain them.

Adaptive management may be valuable to anyone facing substantial uncertainty about the best course of management action, as long as that person has or can obtain the authority to implement a program that leads to learning. In cases where political or other pressures have produced moratoriums or other forms of management "gridlock," such authority may not be easily obtained.[1] Where management interventions are going to proceed (e.g., when delays carry unacceptable social or ecological risks), then much can be gained by treating them as opportunities to learn. In some cases the only way to discover how an action will affect a system is to actually try it (Walters 1986). This is especially true for complex, large-scale systems and effects of cumulative actions.

Although the emphasis in this report and most other literature is the application of adaptive management to natural resources, the approach is equally valuable for other fields. Adaptive management shares much of its theoretical basis with the concepts of continuous improvement in business (Deming 1986; Walton 1986), adaptive control process theory in engineering, and operations research and management (McLain and Lee 1996). Wherever an organized, experimental approach to difficult management questions is needed, adaptive management could be of help.

1.5 Role of Statistics

In forest management, data and mathematical analyses are central to management decision-making, and statistical methods play several important roles. Modern forestry is based to a great extent on statistical descriptions of the characteristics of forests and forest products, such as timber inventories; and on inferences about the expected future conditions of forests and habitat, including growth and yield relations. Statistics are also crucial in understanding how forest resources respond to human and natural perturbations, because they allow us to distinguish "treatment" effects from random and sampling errors.

In adaptive management, statistical methods also play a critical role. Adaptive managers will often want to measure the initial state of the systems they administer, and they will usually need to monitor trends over time that show the system's responses to management policies or practices. In evaluating outcomes, they will want to draw inferences about the causes of any changes that are detected in the

[1] It can be argued that adaptive management is the best way to resolve gridlock based on mistrust of resource managers by public groups or other stakeholders. The same is true where the impasse arises from competing ideas and values, fear of consequences, or other concerns rooted in lack of knowledge of the forest's responses to management.

system to decide how and when to adjust actions in the future or at comparable sites. In all of these applications statistical techniques can provide important insights into both qualitative and quantitative measurements. Statistical analyses allow managers to discern small but important differences in data sets, and to distinguish patterns of correlation and interaction from background variation and sampling errors.

Careful design of management experiments or operational trials is often the first step towards gaining data from which reliable inferences can be drawn. Whenever possible, adaptive management studies should include experimental controls, unbiased sampling and allocation of treatments, and replication of treatments. However, it is important to recognize that the operational scale and setting of adaptive management studies may constrain the level of statistical rigour that can be achieved. It may be impossible, for example, to find multiple areas that are sufficiently homogeneous to serve as replicates of operational-scale treatments. In other cases, it may not be feasible to meet some of the critical assumptions of the classical methods of statistical analysis, including random allocation of treatments, homogeneity of variance, and independence of sample variances. Perhaps even more significant is the fact that "frequentist" statistical methods such as ANOVA and regression analysis are not designed to answer common management questions such as "What is the probability of a 50% increase in windthrow after partial cutting?"

As a result, classical methods will be useful in some adaptive management studies but not in others. When classical methods are not appropriate, a proposed study may still be worthwhile if alternative types of analyses can reveal important insights from the data.

1.6 Structure of the Report

The next eight chapters provide an overview of principles and methods for a wide range of approaches to experimentation and data analysis in operational forestry settings. Beginning with a review of basic concepts of experimental design and classical methods of statistical analysis in Chapter 2, the report then covers other common approaches to studying natural systems ("Studies of Uncontrolled Events" and "Retrospective Studies" in Chapters 3 and 4, respectively). Chapter 5 ("Measurements and Estimates") and Chapter 6 ("Errors of Inference") discuss common mistakes in interpretation of data and statistical results, and suggest ways to avoid them. Chapters 7 and 8 give an overview of Bayesian statistics and decision analysis, both of which are unfamiliar to many resource managers but which have great potential value in adaptive management. Chapter 9 synthesizes the methods discussed in this report and presents a simplified user's guide to the value of different types of information in adaptive management.

Each of the next seven chapters explains what a statistical method can do for a project leader, when it should be used, and what its limitations are. The authors have limited the use of formulas, mathematical notation, and statistical jargon as much as possible, without making the information superficial. Nevertheless, some of the concepts and methods discussed will be unfamiliar and challenging, and some sections may have to be read several times. Most of the technical terms are defined in the glossary at the back of the report. Since this handbook is not comprehensive we encourage readers wanting more detailed information to consult the references listed at the end of each chapter. Most importantly, project leaders should consult throughout their projects with biometricians or experienced researchers to ensure that powerful and cost-efficient methods are used.

Finally, we recognize that there are approaches to generating and analyzing data that this report does not cover. For example, some readers will already be familiar with the methods referred to as "combining information" (Draper et al. 1992) and "meta-analysis" (Fernandez-Duque and Valeggia 1994). Because the focus is on forestry, there is little attention to the quantitative methods for fish and wildlife population analysis that are treated in many adaptive management papers (e.g., Walters 1986). Readers should also remember that this report addresses only one of the many issues that need to be considered if adaptive management is to succeed widely. Greater challenges may lie in social and institutional aspects of implementing adaptive management, such as the risk aversion of some managers (Walters and Holling 1990), inadequate institutional structures and stakeholder participation (McLain and Lee 1996), incomplete or ineffective implementation of the study plan, (C.J. Walters, pers. comm., 1995) uncertain or inadequate funding for monitoring and analyses (McLain and Lee 1996), lack of commitment to reporting (Taylor et al. 1997), and institutional "memory loss" about what has been learned

(Hilborn 1992). These problems too need careful consideration, innovative thinking, and personal commitment to shared learning.

With the stage now set, the remainder of this report presents ideas and methods that should be valuable to anyone who uses or needs to use quantitative analyses in forestry. Throughout, the focus is on designing more powerful and informative management experiments. There is meat, however, for both statisticians and experienced researchers. We hope that the ideas in the following chapters will contribute to more effective, efficient learning in many different situations.

References

Bormann, B.T., P.G. Cunningham, M.H. Brookes, V.W. Manning, and M.W. Collopy. 1994. Adaptive ecosystem management in the Pacific Northwest. U.S. Dep. Agric. For. Serv., Gen. Tech. Rep. PNW-GTR-341.

Covington, W. and P.K. Wagner (technical coordinators). 1996. Conference on adaptive ecosystem restoration and management: restoration of Cordilleran conifer landscapes of North America. Proc. conf. June 6–8, 1996, Flagstaff, Ariz. U.S. Dep. Agric. For. Serv., Gen. Tech. Rep. RM-GTR-278.

Deming, W.E. 1986. Out of the crisis. MIT Center for Advanced Engineering Study, Cambridge, Mass.

Draper, D., D.P. Gaver. Jr., P.K. Goel, J.B. Greenhouse, L.V. Hedges, C.N. Morris, J.R. Tucker, and C.M. Waternaux. 1992. Combining information: Statistical issues and opportunities for research. National Academic Press, Washington, D.C., Contemporary Statistics No. 1.

Fernandez-Duque, E. and C. Valeggia. 1994. Meta-analysis: a valuable tool in conservation research. Cons. Biol. 8:555–61.

Gunderson, L.H., C.S. Holling, and S.S. Light (editors). 1995. Barriers and bridges to the renewal of ecosystems and institutions. Columbia Univ. Press, New York, N.Y.

Hilborn, R. 1987. Living with uncertainty in resource management. N. Am. J. Fish. Manage. 7:1–5.

______. 1992. Can fisheries agencies learn from experience? Fisheries 17:6–14.

Holling, C.S. (editor). 1978. Adaptive environmental assessment and management. J. Wiley, London, U.K.

Lancia, R.A., C.E. Braun, M.W. Collopy, R.D. Dueser, J.G. Kie, C.J. Martinka, J.D. Nichols, T.D. Nudds, W.R. Porath, and N.G. Tilghmann. 1996. ARM! for the future: adaptive resource management in the wildlife profession. Wildl. Soc.Bull. 24:436–42.

Lee, K.N. 1993. Compass and gyroscope: integrating science and politics for the environment. Island Press, Washington, D.C.

McLain, R.J. and R.G. Lee. 1996. Adaptive management: promises and pitfalls. Environ. Manage. 20:437–48.

MacDonald, G.B., R. Arnup, and R.K. Jones. 1997. Adaptive forest management in Ontario: a literature review and strategic analysis. Ontario Min. Nat. Resour., For. Res. Info. Pap. 139.

Namkoong, G. 1997. A gene conservation plan for loblolly pine. Can. J. For. Res. 27:433–7.

Nyberg, J.B. and B. Taylor. 1995. Applying adaptive management in British Columbia's forests. *In* Proc. FAO/ECE/ILO International Forestry Seminar, Prince George, B.C., Sept. 9–15, 1995, pp. 239–45. Can. For. Serv., Prince George, B.C.

Schmiegelow, F.K.A. and S.J. Hannon. 1993. Adaptive management, adaptive science and the effects of forest fragmentation on boreal birds in northern Alberta. Trans. N. Am. Wildl. Nat. Resour. Conf. 58:584–97.

Taylor, B., L. Kremsater, and R. Ellis. 1997. Adaptive management of forests in British Columbia. B.C. Min. For., For. Practices Br., Victoria, B.C.

Walters, C.J. 1986. Adaptive management of renewable resources. McGraw-Hill, New York, N.Y.

Walters, C.J. and C.S. Holling. 1990. Large-scale management experiments and learning by doing. Ecology 71:2060–8.

Walton, M. 1986. The Deming management method. Perigee Books, New York, N.Y.

2 DESIGN OF EXPERIMENTS

AMANDA F. LINNELL NEMEC

Abstract

Experimentation is essential for making well-informed decisions about the management of complex forest ecosystems. Although experiments in forest management are generally more complicated than the typical research experiment, many issues, such as protection against bias, repeatability of results, efficient use of resources, and quantification of uncertainty, are the same. Therefore, managers, as well as researchers, can benefit from a good understanding of the principles of sound experimental design. This chapter reviews some fundamental concepts of experimental design and the assumptions that provide a basis for classical statistical inference. The importance of clear objectives and the need for replication, randomization, and blocking are emphasized. Practical limitations of the classical approach, as it applies to the design of adaptive management experiments, are discussed briefly.

2.1 Introduction

Successful management of our forests is a dynamic process in which current programs are continually monitored and adapted as new information becomes available and policies change. Because the outcome of decisions is always uncertain, managers often experiment with new strategies to help determine the best course of action. Although such tests do not necessarily conform to the standards of a strictly controlled research experiment, many issues, such as elimination of bias, repeatability of results, efficient use of resources, and quantification of uncertainty, are the same. For this reason, managers, as well as researchers, can benefit from a good understanding of the principles of sound experimental design. This chapter reviews the theory of classical experimental design and the assumptions that provide a basis for statistical inference from experimental data. Application of traditional theory to the design of adaptive management experiments is considered.

The literature on the design of experiments is vast. Classical books, such as *The Design of Experiments* by Fisher (1935), *The Design and Analysis of Experiments* by Kempthorne (1952), *Experimental Designs* by Cochran and Cox (1957), and *Planning of Experiments* by Cox (1958), remain valuable sources of guidance and are recommended reading. Since the early work of Fisher, the number of books and papers on experimental design has exploded, as indicated by the list of more than 200 references assembled over 10 years ago by Steinberg and Hunter (1984). During the last 10 years, improvements in computer technology have encouraged statisticians to improve and expand their efforts even more (e.g., refer to Atkinson 1982, 1988, 1996; Bates et al. 1996), so the trend persists. Unfortunately, much of the recent work is inaccessible to managers because the language is overly technical and difficult to understand—a concern expressed by Pike (1984) and several others who discuss the review by Steinberg and Hunter (1984). Moreover, there is usually a long delay before new methods are accepted and incorporated into popular statistical software packages. Therefore, readers seeking practical information on the design and analysis of experiments are advised to consult general textbooks (e.g., John 1971; Box et al. 1978; Steel and Torrie 1980; Mead 1988; Montgomery 1991; Milliken and Johnson 1989, 1992) and papers written specifically for managers or applied scientists (e.g., Stafford 1985; Penner 1989).

2.2 Definitions

An experiment is the manipulation of a population or system (e.g., partial cutting of a forested area) as a means of gathering information. The source of experimental material (e.g., trees, vegetation, wildlife) is called the experimental population. In an ideal experiment, the experimental population is the same as the target population, or forest ecosystem, to which the results are eventually to be applied. In practice, the two populations necessarily differ in time and perhaps space or scale as well.

In a typical research experiment, the experimental population might consist of: the trees in a nursery; a collection of relatively small plots of land, which are distributed over one or more geographic areas; or any other set of entities. Each tree, plot, or entity is an experimental unit that receives one of several treatments (e.g., levels of fertilizer, partial cutting or no cutting). These units are generally too large to measure in their entirety and are subdivided into a set of smaller sampling units (e.g., branches or subplots) from which a suitable subset is selected for measurement. To monitor trends or other effects of time, measurements are repeated at suitable intervals over a period of months or years.

The results of research experiments are of limited value for making inferences about the effects of treatments or actions applied to management units. Management units (e.g., a forest stand or polygon, a mapsheet, or a timber supply area) are areas of forest that are convenient for administration or cost-effective operations, and therefore are considerably larger than research plots. The impact of disturbances created by large-scale operations in these units cannot, in general, be deduced from small research plots, where effects such as fragmentation, soil erosion, and changes in vegetation or water quality might not be evident. Proponents of adaptive management (e.g., Walters 1986; Walters and Holling 1990; Taylor et al. 1997) argue that successful management of complex biological systems requires full-scale testing. These experiments, which are known as adaptive management experiments, are used to test entire management plans, with the management unit serving as the experimental unit as illustrated in Figure 2.1. In an adaptive management experiment, one or more systems are monitored regularly over time and decisions about treatments or other interventions are made as the experiment progresses. Because management units are large and complex, they must be broken down into suitable sampling units for observation and evaluation. In this respect, adaptive management experiments resemble research experiments, although the number and type of sampling units might differ.

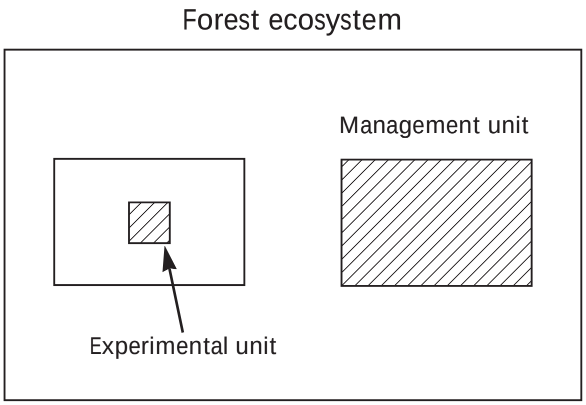


FIGURE 2.1 *Relationship between the study units in a typical research experiment (left) and an adaptive management experiment (right).*

2.3 Objectives

The first requirement in any experiment is a clear statement of the goals. Paraphrasing Box et al. (1978, p. 15), the purpose of an experiment must be clear and agreed upon by all interested parties; there must be agreement on what criteria will determine whether the objectives have been accomplished; and, finally, in the event that the objectives change during the course of the experiment, an arrangement must be made by which all interested parties are informed of the change, and agreement on the revisions can be reached. The importance of these points cannot be overemphasized. Without clear objectives, the outcome of an experiment is predictable: ambiguous results and wasted time, money, and valuable (possibly irreplaceable) resources. Unnecessary waste is always unacceptable. However, when large management units are involved, the costs can be devastating.

Defining the objectives of an experiment requires careful consideration of the components that make up the system under study, the forces that drive the system, and the best means of extracting information about both. In a small-scale research experiment, attention might reasonably be restricted to relatively simple questions—for instance, how does tree growth differ under various controlled conditions? The objectives of adaptive management experiments typically concern more complicated issues—such as, how is "biodiversity" affected by forest practices? In both cases, general scientific concepts (e.g., tree growth and biodiversity) must be stated in terms of well-defined, measurable quantities (e.g., height or diameter increment, number of species). These quantities provide a concrete basis for planning experiments and for analyzing the results.

The objectives of an experiment are often posed as hypotheses to be tested or parameters to be estimated. Special care must be taken to ensure that the hypotheses are sensible and that the parameters are useful for making decisions. In classical hypothesis testing, a so-called null hypothesis is retained unless there is convincing evidence to the contrary. Based on the outcome of the experiment, the null hypothesis is either retained or rejected in favour of a specific alternative hypothesis. (See Anderson, this volume, Chap. 6, for a discussion of the associated errors of inference.) The null and alternative hypotheses should be defined so that both outcomes (i.e., acceptance or rejection of the null hypothesis) represent reasonable and informative conclusions that would justify the cost of the experiment. Implausible null hypotheses should be rejected at the outset and replaced with something more relevant. For instance, no sensible person would ever accept the hypothesis that partial cutting has no impact on forest ecosystems; a more reasonable hypothesis is that partial cutting does not reduce a key parameter (e.g., the number of bird species) below some critical value. The alternative hypothesis is equally, if not more, important than the null hypothesis. An experiment that is designed to detect one type of departure from a null hypothesis might be completely ineffective in detecting other types of deviations. For example, an experiment to evaluate short-term impacts of partial cutting will generally provide little information about long-term effects.

2.4 Principles of Experimental Design

An experimental design is a detailed plan describing all aspects of an experiment (see Bergerud 1989b). Most experimental designs include a sampling design, which describes the nature of the sampling units, the number of sampling units, the method of selection, and the variables to be measured. The experimental design depends on the purpose of the experiment and thus "the entire reasoning process by which the experimenter really hopes to link the facts he wishes to learn from the experiment with those of which he is already reasonably certain" (Mandel 1964, p. 2). Experimental designs are characterized by three main components: (1) the factors and factor levels to be investigated, (2) the amount and type of replication, and (3) the method of randomization, including any blocking. Each of these elements should be considered carefully to assure that all data pertaining to the objectives are collected (and can be analyzed) and that the data be collected in the most efficient way (i.e., optimum results for a minimum cost).

2.4.1 Experimental factors

Experimental factor means any treatment[1] or variable (e.g., harvesting method, species composition, age) that is controlled in an experiment, either by physically applying a treatment to an experimental unit or by deliberately selecting a unit with a particular characteristic. A covariate may be any other variable that is measured but not influenced by the experiment. Experiments are distinguished from observational studies by the investigator's ability to control the experimental conditions. (Refer to Eberhardt and Thomas 1991, and Schwarz, this volume, Chap. 3, for a discussion of the differences among experiments, observational studies, and various other types of study.) This control helps justify inferences about cause and effect. For example, to determine the best method of partial cutting to minimize the impact on wildlife, various levels of volume removal might be tested. This approach would probably be more informative than simply measuring volume in existing partial cuts and observing a correlation with number of animals. Correlation proves little about cause and effect because both variables might be correlated with a third variable (e.g., elevation). An interesting discussion of inferences about cause and effect can be found in Holland (1986).

Usually many types and levels of factors might be investigated in an experiment. These must be selected carefully to meet the study objectives. A simple experiment might have only one relevant factor (e.g., method of logging), which has relatively few predetermined values or levels (e.g., clearcutting, partial cutting, or no cutting). An experimental design that involves only one such factor is called a one-way design with a fixed factor (effect). A design with two such factors is called a two-way design, a design with three factors is a three-way design, and so on. In some applications, the factor levels are not known ahead of time (although the number of levels is fixed) but are randomly selected from a well-defined population of levels. Such factors are known as random factors (effects). For instance, a manager might want to compare bird diversities in a random sample of 20 white pine stands from a particular site series. Here stand is a random factor with 20 levels. Factors with a continuum of possible values (e.g., age, diameter), which commonly occur in regression designs, are called continuous factors or simply variables.

Two or more factors can occur in various combinations. If an experiment includes every level of one factor in combination with every level of the other factor(s) then the factors are crossed and the design is called a factorial design. The advantages of factorial designs over designs that vary one factor at a time are twofold: efficient replication and estimates of interactions. In some cases, the set of possible levels of one factor depends on the level of another, in which case the former is said to be nested in, rather than crossed

[1] The term "treatment" will hereafter refer to a particular set of conditions, an action, or an entire management strategy.

with, the latter. The (relative) moisture level of a stand might, for example, be classified differently depending on the subzone in which the stand is located. Most well-designed experiments include a control or standard by which the effectiveness or impact of treatments can be judged. In a one-way design, the control might be one level of the treatment (e.g., clearcut, partial cut, and an uncut control), while a two-way factorial design might include a separate control (e.g., clearcut with and without mechanical site preparation, partial cutting with and without site preparation, and an uncut control with no site preparation). Bergerud (1989a) describes the analysis of the latter type of design.

The overall effect of a factor—that is, the average effect for all levels of the other factors—is called its main effect. If the effect of one factor depends on the level of one or more other factors then there is an interaction among the factors. A certain amount of fertilizer might, for example, increase the height of one species but have a smaller, or even opposite, effect on another species; in this case, there is an interaction between species and fertilizer. Factorial designs allow an investigator to study many interactions in the same experiment. However, if some combinations are omitted for any reason then certain main effects or interactions may be inseparable from, or confounded with, others. Thus the magnitude of confounded effects cannot be estimated. In such circumstances, the experimenter must be sure that the confounded effects are of little or no interest, or can be assumed to be negligible.

2.4.2 Replication

Replication, a standard means of validating scientific findings, is a cornerstone of the theory of experimental design as laid down by Fisher (1935). Experimental conditions are replicated when the same combination of factors occurs in more than one experimental unit. This replication provides an estimate of the experimental error, which is any variation that cannot be explained by the experimental factors (e.g., sampling or measurement error, and natural variation among the experimental units). In the absence of replication, there is no way, without appealing to nonstatistical arguments, to assess the importance of observed differences between the experimental units.

The type of replication should be consistent with the objectives. The experimental units should, therefore, be as similar as possible to the elements of the target population. For instance, test sites should be selected from the same geographic area as the target population and treatment plots should be comparable in type and size to management units. Experimentation at a single site limits the applicability of results to a small geographic region, while failure to identify the appropriate unit of replication results in pseudoreplication (Hurlbert 1984; Bergerud 1988, 1991) and possibly erroneous conclusions about the nature of an effect.

In addition to selecting the size and type of replicate, the optimum number of experimental units (and sampling units) must be determined. Sample sizes should be large enough that definitive conclusions about the size of an effect or the validity of a hypothesis can be made (i.e., parameter estimates must be sufficiently precise and tests of hypotheses must be decisive). Simple sample size calculations provide an estimate of the minimum number of replicates and sampling units needed to meet this goal (see Nemec 1991; Bergerud 1992). If the required sample size is unrealistic, because of the high cost of treatment or sampling, then the objectives and design of the experiment should be re-evaluated to determine whether an observational or retrospective study (see Schwarz, this volume, Chap. 3, and Smith, this volume, Chap. 4) might be more cost effective.

2.4.3 Randomization and blocking

Randomization occurs when treatments are randomly assigned to the experimental units. Like replication, randomization is an essential element of good design. It helps minimize the risk of bias by ensuring that all unmeasured factors (e.g., soil moisture, nutrients, prevalence of root disease) are more or less evenly distributed among the treatments, or "randomized out" of the experiment. Thus each treatment is expected to be assigned to an approximately equal number of wet and dry sites, nutrient-rich and nutrient-poor sites, sites with low and high levels of root disease, and so on.

Random assignment of treatments can be accomplished in a variety of ways. In a completely randomized design, treatments are assigned by picking units at random from the experimental population. For instance, if there were two treatments and 20 experimental units (10 replicates per treatment) then the units might be numbered from 1 to 20, with 10 numbers picked at random to determine which units receive the first treatment and the remainder receiving the second. Sometimes experimental units have a variety of origins (e.g., different geographic regions) and therefore exhibit considerable inherent variability. In such situations, a substantial reduction in the experimental error can often be achieved by separating the units into homogeneous groups, or blocks, according to origin, or some other factor. Treatments are then randomly assigned within each block. This is a form of restricted randomization because each treatment is constrained to occur a fixed number of times in each block. In a completely randomized design, the randomization is unrestricted, which might result in a very uneven distribution of origins or other attributes among the treatment groups. Many other designs with some form of restricted randomization exist, including split-plot designs (which impose constraints on the assignment of treatments to two types of experimental units: main plots and subplots), Latin squares, and lattice designs. Refer to Cochran and Cox (1957) and Anderson and McLean (1974) for more information about these and other designs.

Random sampling is another way of avoiding bias when factors, such as species or age, cannot be assigned. Experimental units should be randomly chosen from the experimental population and sampling units should likewise be selected at random from the experimental units. Random samples, unlike haphazard samples or judgement samples (samples judged to be "representative" by an expert who makes the selection), have known statistical properties. This allows the precision of a result to be estimated from the sample—something that is not possible with non-random sampling. For further discussion of the topic, see Deming (1960, pp. 30–33) and Schwarz (this volume, Chap. 3).

Randomization and random sampling are closely related ideas leading, in many cases, to the same mathematical models and equivalent data analyses (Feinberg and Tanur 1987; Smith and Sugden 1988). For instance, randomization with blocking is analogous to stratified random sampling, and split-plot designs are comparable to cluster sampling. Despite the parallels, randomization is traditionally discussed in connection with experimental design, while issues relating to sampling design (e.g., type and number of sampling units, method of sampling) are reserved for discussions of observational studies. For more information on the latter subject refer to Schwarz (this volume, Chap. 3).

Randomization is one of the simplest ideals to achieve in both small- and large-scale experiments. The process of randomly assigning treatments to units is the same regardless of the size or nature of the unit—tree, research plot, or stand. All that is required is a list of units and a random number generator. Despite its importance in eliminating bias and its ease of application, randomization is sometimes resisted on grounds that "environmental" or "logistical" constraints prohibit its use, or that it is "impractical." Although it might be tempting to assign certain treatments (or controls) to stands that are most difficult to harvest, or to stands that are most visible from neighbouring communities, such practices are likely to influence the outcome and thus invalidate any inferences about cause and effect (refer to Section 2.5). If randomization of treatments is really not practical or possible then the whole purpose of the experiment should be reconsidered.

Random selection of experimental units and sampling units can be more difficult to achieve than random assignment of treatments. For instance, how can a random sample of 25 trees be selected from a stand without compiling a complete list of trees and their locations? Likewise, how can a random sample of needles be selected from a tree? Various ingenious solutions have been proposed, such as the use of random bearings and random distances to locate sample trees and randomized-branch sampling to sample individual trees (Gregoire et al. 1995). Thus both randomization and random sampling (or a close approximation) are feasible in most experiments.

2.4.4 Other considerations

Replication, randomization, and blocking are usually recognized as the most important principles of good design. These have been expanded over the years (see Atkinson 1982, 1988; Federer 1984; Steinberg and Hunter 1984) to include such other features as orthogonality (which implies that all main effects and interactions of the experimental factors can be estimated), balance (equal sample sizes), efficiency (minimum cost), and, in recent years, optimality (minimum variance). Efficiency and optimality, which help ensure maximum precision for a fixed sample size, are likely to remain important as long as money and resources are scarce. As technology continues to improve, the emphasis on orthogonality and other computational issues is expected to diminish. Mead (1990) has, for example, suggested that the need for orthogonality, which excludes a large class of potentially useful designs, primarily because of numerical complexity, should be re-evaluated in the light of present computing power. As the role of experiments continues to evolve, new criteria and principles will undoubtedly emerge (see Section 2.6).

Successful experimentation depends on more than the choice of factors and the use of appropriate randomization, replication, and blocking. Many practical details must also be considered. All pertinent measurements of the experimental (sampling) units must be identified, appropriate field procedures and data collection forms must be developed, and provision must be made for adequate supervision of the data collection. In large-scale studies, coordination and optimization of procedures are especially important. For a checklist of these and other aspects of the planning and execution of a study, refer to Sit (1993).

2.5 Statistical Inference for Experiments

Statistical inference and experimental design are closely linked. Design determines the kind of statistical inferences that are possible, while consideration of the proposed method of analysis almost always influences design (e.g., sample-size calculations depend on the hypothesis testing procedure or estimation method that will be used). In fact, failure to contemplate how the data will be analyzed invariably results in a poor design.

All statistical inferences are based on a set of assumptions that links the data to the experimental population via the design. Thus inferences are necessarily confined to the experimental population. Because the relationship between experimental and target populations is unknown, extrapolation to the latter is more or less conjecture and should be viewed with caution, particularly when the gap between experimental and target populations is large. The validity of statistical inference in an experimental setting is discussed more fully by Deming (1953, 1975), Box et al. (1978, Chap. 1), and Hahn and Meeker (1993).

Analysis of variance (ANOVA) methods are commonly applied to experimental data, so much so that discussions of experimental design are often more about ANOVA than fundamental issues of design. Both are concerned with sources of variation and the estimation of experimental error. The basic premise of ANOVA is that the observed variability in experimental data can be attributed to a finite number of identifiable sources, including factors under the control of the investigator, uncontrolled experimental errors, and various interactions. A simple, one-way, fixed-effects design has a single experimental factor (e.g., harvesting method) and a single source of experimental error (e.g., inherent variability in the stands that are treated). A split-plot design has two sources of experimental error: variation among main plots and variation among subplots. Proper data analysis is possible only when the identification and classification of the factors by type (e.g., fixed or random, nested or crossed) is consistent with the design. Any restrictions on the randomization, which are imposed by design (e.g., blocking), must be duly noted by the inclusion of appropriate error terms (a theme emphasized by Anderson and McLean 1974).

The purpose of an ANOVA is to make inferences about effects attributable to experimental factors (refer to Sit 1995 for more information), while taking into account uncertainty caused by errors. To do this the data are assumed to be generated by a probability model that includes all the relevant effects and has three key assumptions: (1) selection of the experimental units and assignment of treatments are independent of the response variables of interest, (2) all random effects and experimental errors are mutually independent, and (3) random effects and experimental errors attributable to a common source are identically distributed as (normal) random variables with a mean of zero. The first assumption is often not stated explicitly but is crucial for making causal inferences (refer to Holland 1986 for a mathematical explanation of this fact). If the variables of interest somehow influence, either directly or indirectly, which experimental units receive a particular treatment, then the potential for bias is clear (e.g., assignment of treatment on the basis on slope, aspect, density of trees, or any other variable that might affect tree height precludes inferences about the effect of treatment on height). Randomization and random sampling are the only effective means of eliminating this source of bias.

Independence among errors and random effects is another important assumption of ANOVA. There are two common departures from this condition: temporal and spatial autocorrelation. In forestry studies, correlation among repeated measurements of the same experimental unit (temporal autocorrelation) and correlation among units in close proximity (spatial autocorrelation) are likely occurrences. The possibility of autocorrelation can sometimes be avoided by an appropriate choice of design (e.g., by ensuring that the sampling units are far apart) or it can be taken into account by adopting a suitable spatial or time-series probability model (refer to Nemec 1996)

Formal statistical inference based on ANOVA methods (e.g., *F*-tests, computation of confidence intervals) requires knowledge of the distributional properties of the random effects and experimental errors. The usual assumption is that all random effects or experimental errors associated with a specific source have the same distribution, which is generally assumed to be normal with a mean of zero and a homogeneous variance. The impact of departures from these assumptions varies depending on the type of departure and on the sample size. Minor deviations from normality have little impact on inferences about fixed effects, especially when the sample sizes are large. Tests for random effects tend to be more seriously affected. Heterogeneous error variances can distort P-values, although the degree of distortion depends on the range of variances. Serious departures from normality or homogeneity can sometimes be identified and corrected; some suggestions are outlined by Snedecor and Cochran (1967, Chap. 11). Alternatively, robust or nonparametric methods can be used. Failure to adjust for lack of independence (autocorrelation) tends to exaggerate effects when the correlation is positive, or understate effects when the correlation is negative.

2.6 Advantages of Classical Experimental Design

The "controlled experiment is in common wisdom the most decisive of tests" (Susser 1994, p. 831). By controlling both the levels and combinations of factors applied to the experimental material, the investigator can make much stronger inferences about causal relationships and interactions than would otherwise be possible. Replication, randomization, and blocking serve three important purposes: elimination of systematic error (i.e., bias), quantification of uncertainty, and reduction of uncontrolled experimental error. Because these three goals are relevant to any scientific investigation, many parallels can be found between experimental design and the design of a survey or an observational study (see Schwarz, this volume, Chap. 3). Two examples mentioned previously are randomization and random sampling to eliminate bias, and blocking and stratification to improve precision; refer to Feinberg and Tanur (1987) and Smith and Sugden (1988) for a discussion of these and other similarities and differences.

Classical experimental design also provides a framework for ANOVA, analysis of covariance, regression methods, and other types of statistical analysis. This framework allows investigators to estimate components of variation, and, beyond that, to make formal statistical inferences about their significance. Hypothesis testing has been emphasized in the past but confidence intervals, which are often more informative because they provide an estimate of the size of an effect, are just as easily constructed. The usefulness of ANOVA—both the underlying models and the components-of-variation approach—cannot be disputed. The models are very versatile. Simple forms of temporal or spatial autocorrelation can, for instance, be accommodated by repeated-measures or split-plot models. However, the usual ANOVA assumptions (e.g., normality) are untenable in some situations (e.g., if the data are discrete). This is a limitation of the probability model, not of the classical approach to design. Although ANOVA models and experimental design are closely linked, the two should not be equated. Other models (e.g., log-linear models for categorial data, nonparametric models) and methods of analysis (e.g., regression analysis) might be applicable when ANOVA models are not. In addition, new models and methods continue to be developed, many of them for classical designs.

2.7 Experimental Design for Adaptive Management

An experiment is conducted to answer one or more questions. In a research setting, the questions tend to be relatively simple. Does a new method of storage promote better short-term survival of seedlings than an established method? Is there any difference in the growth of two species of seedlings, 5 years after planting? Are planted trees more susceptible to root disease than naturally regenerated trees? The questions confronting managers are likely to be considerably more challenging, involving such complex issues as maximization of sustainable yield, avoidance of unnecessary risk, economic efficiency, and economic stability (see Walters 1986, Chap. 2). To limit these problems, thereby making them more amenable to solution by experimentation, a manager must: (1) identify key factors for analysis: what are the main factors that distinguish strategies? which factors can be controlled? (2) consider timing: over what time period and how frequently should the system be monitored? (3) consider spatial scale: what are the management units? and (4) identify quantities of interest: how are quality of results, costs, and benefits measured? For practical advice on "bounding problems for analysis," refer to Chapter 1 of Walters (1986).

After the objectives have been determined, the next step is to define the target population and to take an inventory of the management units available for experimentation. If the target population is a unique ecosystem managed as a single unit (e.g., a particular area of old-growth forest) then replication is inapplicable because there can be only one response to a particular management strategy. On the other hand, when groups of units are sufficiently similar that they can be managed according to a common strategy (e.g., stands with comparable ages and species compositions), replication is desirable to determine the range of possible responses to that strategy. However, even when replication is theoretically possible, it might not be practical because of the high costs of large-scale experiments, time constraints, or limited resources.

Replication is a decisive issue. Its effect on the design and analysis of adaptive management experiments is illustrated in Figure 2.2. If replication of a treatment or management strategy is possible then classical methods (Sections 2.4 and 2.5) can be useful. Moreover, even if treatment replication is impossible or impractical, adherence to traditional principles can help ensure that the sampling design is sound. Thus, randomization, replication, and blocking (random sampling and stratification) are effective means of avoiding bias and reducing error in both replicated and non-replicated experiments.

Adaptive management requires a suitable model for predicting transitions of a system from one state to another, and a set of rules for deciding the best action at any given time. In the case of non-replicated experiments (right side of Figure 2.2), various analytical methods have been developed (see Walters 1986, Chap. 4–9). These methods are based on the theory of stochastic processes, Bayesian statistics (Bergerud and Reed, this volume, Chap. 7), and decision theory (Peterman and Peters, this volume, Chap. 8). When data arise from replicated systems (left side of Figure 2.2), the problem is considerably more complicated. Responses of individual systems and the overall response of systems managed under the same plan (i.e., replicates) must be considered (see Walters 1986, Chap. 10). This problem has no simple solution because the link between classical methods (e.g., ANOVA) for the analysis of replicated designs and decision analysis for nonreplicated management strategies is not well developed. Meta-analysis (see Mann 1990 for an interesting and nontechnical discussion of meta-analysis) or alternative methods for integrating the results from several experiments might be useful in such situations, although a piece-meal analysis of large, complex, and dynamic systems has obvious drawbacks.

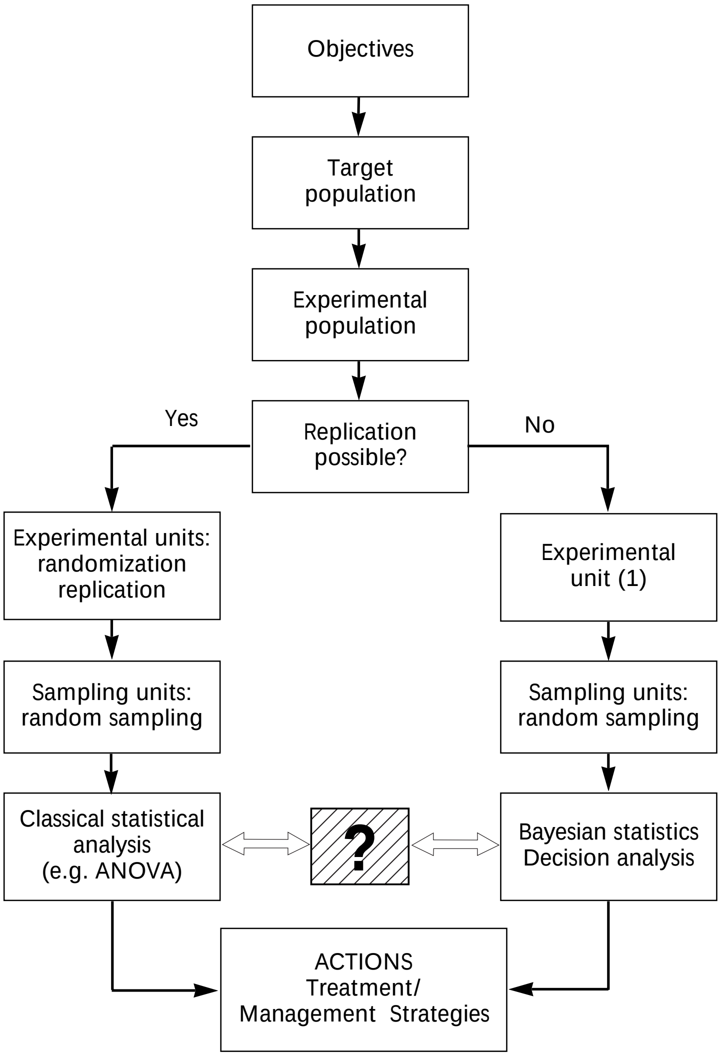

Objectives
Target population
Experimental population
Yes
Replication possible?
No
Experimental units: randomization replication
Experimental unit (1)
Sampling units: random sampling
Sampling units: random sampling
Classical statistical analysis (e.g. ANOVA)
?
Bayesian statistics Decision analysis
ACTIONS Treatment/ Management Strategies

FIGURE 2.2 *Design and analysis of an adaptive management experiment.*

2.8 Summary

Sound experimental design is essential for adaptive management of valuable forest resources. Adherence to the principles of randomization, replication, and blocking helps to ensure that an experiment meets the basic requirements for success (Cox 1958): absence of systematic error, precision, validity for an appropriate range of conditions, simplicity, and an estimate of uncertainty. Failure to consider these issues leads to unnecessary waste and, in the worst case, bad decisions resulting in serious damage to sensitive forest ecosystems. Development and adoption of adaptive methods for carrying out large-scale experiments has been slow, due partly to the barriers created by excessively technical language, a lack of analytical tools, and a limited number of successful applications to serve as models. Overcoming these obstacles by continuing efforts to educate and to inform (e.g., Biometrics Information series published by Research Branch, B.C. Ministry of Forests) can only help to improve matters in the future.

References

Anderson, J.L. [n.d.]. Errors of inference. This volume.

Anderson, V.L. and R.A. McLean. 1974. Design of experiments: a realistic approach. Marcel Dekker, New York, N.Y.

Atkinson, A.C. 1982. Developments in the design of experiments. Intern. Statist. Rev. 50:161–77.

———. 1988. Recent developments in the methods of optimum and related experimental designs. Intern. Statist. Rev. 56:99–115.

———. 1996. Usefulness of optimum experimental designs. J. Royal Statist. Soc., B, 58:59–76.

Bates, R.A., R.J. Buck, E. Riccomgno, and H.P. Wynn. 1996. Experimental design and observation for large systems. J. Royal Statist. Soc., B, 58:77–94.

Bergerud, W. 1988. Understanding replication and pseudoreplication. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 5.

______. 1989a. ANOVA: Factorial designs with a separate control. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 14.

______. 1989b. What is the design? B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 17.

______. 1991. When are blocks pseudoreplicates? B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 34.

______. 1992. A general description of hypothesis testing and power analysis. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 37.

Bergerud, W.A. and W.I. Reed. [n.d.] Bayesian statistical methods. This volume.

Box, G.E.P., W.G. Hunter, and J.S. Hunter. 1978. Statistics for experimenters: an introduction to design, data analysis, and model building. J. Wiley, New York, N.Y.

Cochran, W.G. and G.M. Cox. 1957. Experimental designs. 2nd ed. J. Wiley, New York, N.Y.

Cox, D.R. 1958. Planning of Experiments. J. Wiley, New York, N.Y.

Deming, W.E. 1953. On the distinction between enumerative and analytic surveys. J. Am. Statist. Assoc. 48:244–55.

______. 1960. Sample design in business research. J. Wiley, New York, N.Y.

______. 1975. On probability as a basis for action. Am. Statist. 29:146–52.

Eberhart, L.L. and J.M. Thomas. 1991. Designing environmental field studies. Ecol. Monogr. 61:53–73.

Federer, W.T. 1984. Principles of statistical design with special reference to experiment and treatment design. *In* Statistics: An Appraisal. Proc. 50th Anniv. Conf. Iowa State Statist. Lab. H.A. David and H.T. David (editors), Iowa State Press, pp. 77–105.

Feinberg, S.E and J.M. Tanur. 1987. Experimental and sampling structures: Parallels diverging and meeting. Intern. Statist. Rev. 55:75–96.

Fisher, R.A. 1935. The design of experiments. Hafner, New York, N.Y. Reprint 1971.

Gregoire, T.G., H.T. Valentine, and G.M. Furnival. 1995. Sampling methods to estimate foliage and other characteristics of individual trees. Ecology 76:1181–94.

Hahn, G.J. and W.Q. Meeker. 1993. Assumptions for statistical inference. Am. Statist. 47:1–11.

Holland, P.W. 1986. Statistics and causal inference. J. Am. Statist. Assoc. 81:945–60.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211.

John, P.W.M. 1971. Statistical design and analysis of experiments. McMillan, New York, N.Y.

Kempthorne, O. 1952. The design and analysis of experiments. Krieger, Malabar, Fla.

Mandel, J. 1964. The statistical analysis of experimental data, Dover, Mineola, N.Y.

Mann, C. 1990. Meta-analysis in the breech. Science 249:476–80.

Mead, R. 1988. The design of experiments, statistical principles for practical applications. Cambridge Univ. Press, Cambridge, U.K.

______. 1990. The non-orthogonal design of experiments. J. Royal Statist. Soc., A, 153:151–201.

Milliken, G. and D.E. Johnson. 1989. Analysis of messy data. Vol. 2: nonreplicated experiments. Van Nostrand Reinhold, New York, N.Y.

______. 1992. Analysis of messy data. Vol. 1: designed experiments. Wadsworth, Belmont, Calif.

Montgomery, D.C. 1991. Design and analysis of experiments. 3rd ed. J. Wiley, New York, N.Y.

Nemec, A.F.L. 1991. Power analysis handbook for the design and analysis of forestry trials. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Handb. No. 2.

______. 1996. Analysis of repeated measures and time series: an introduction with forestry examples. B.C. Min. For., Res. Br., Victoria, B.C. Work. Pap. 15/1996. Biometrics Inf. Handb. No. 6.

Penner, M. 1989. Optimal design with variable cost and precision requirements. Can. J. For. 19:1591–7.

Peterman, R.M. and C. Peters. [n.d.] Decision analysis: taking uncertainties into account in forest resource management. This volume.

Pike, D.J. 1984. Discussion (of paper by Steinberg and Hunter 1984). Technometrics 26:105–9.

Schwarz, C.J. [n.d]. Studies of uncontrolled events. This volume.

Sit, V. 1993. What do we look for in a working plan? B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 44.

______. 1995. Analyzing ANOVA designs. B.C. Min. For., Res. Br., Victoria, B.C. Work. Pap. 07/1995. Biometrics Inf. Handb. No. 5.

Smith, G.J. [n.d.] Retrospective studies. This volume.

Smith, T.M.F. and R.A. Sugden. 1988. Sampling and assignment mechanisms in experiments, surveys, and observational studies. Int. Statist. Rev. 58:165–80.

Snedecor, G.W. and W.G. Cochran. 1967. Statistical methods. 6th ed. Iowa State Univ. Press, Ames, Iowa.

Stafford, S.G. 1985. A statistics primer for foresters. J. For. 83:148–57.

Steel, R.G. and J.H. Torrie. 1980. Principles and procedures of statistics: a biometrical approach. 2nd ed. McGraw-Hill, New York, N.Y.

Steinberg, D.M. and W.G. Hunter. 1984. Experimental design: review and comment. Technometrics 26:71–130.

Susser, M. 1994. The logic in ecologic: II. the logic of design. Am. J. Public Health 84:830–5.

Taylor, B., L. Kremsater, and R. Ellis. 1997. Adaptive management of forests in British Columbia. B.C. Min. For., For. Practices Br., Victoria, B.C.

Walters, C. 1986. Adaptive management of renewable resources. MacMillan, New York, N.Y.

Walters, C.J. and C.S. Holling. 1990. Large-scale experiments and learning by doing. Ecology 71:2060–8.

3 STUDIES OF UNCONTROLLED EVENTS

CARL J. SCHWARZ

Abstract

The rationale for carefully planned experiments is well documented. A well-designed experiment will have a high probability of detecting important, biologically meaningful differences among the experimental groups. Causal relationships can be inferred because the experimental factors have been manipulated and randomly assigned.

In many cases, controlled experiments are impractical or too expensive, and surveys of existing ecological populations are performed, even though the resulting inferences will be weaker than those obtained through controlled experimentation. Consequently, nonexperimental studies, or passive adaptive management, is primarily a tool for generating hypotheses to be tested by careful experimentation.

Despite the weaker inferences from nonexperimental studies, the same attention must be paid to the proper design of a survey so that the conclusions are not tainted by inadvertent biases. This paper will review several of the standard nonexperimental studies by presenting an overview of the study protocol, the conclusions that can be reached, and the potential problems that can occur.

3.1 Introduction

The rationale for carefully planned experiments in ecology is well documented (Hurlbert 1984). A well-designed experiment will have a high probability of detecting important, biologically meaningful differences among the experimental groups. Furthermore, because the manager directly manipulated the experimental factors and randomly assigned the experimental units to the particular combination of experimental factors, the manager can infer a causal relationship between the experimental factors and the response variable. The manager who takes a similar approach and practices active adaptive management can make the strongest possible inferences about the role of the experimental factors.

In many cases, controlled experiments are impractical or too expensive, and surveys of existing ecological populations are performed, even though the resulting inferences will be weaker than those obtainable through controlled experimentation.

Consequently, nonexperimental surveys, or passive adaptive management, leads to conclusions that are primarily a tool for generating hypotheses eventually to be tested by careful and more efficient experimentation.

For example, observation surveys of existing lakes showed that the more acidic lakes tended to have fewer fish. An alternative explanation that could "explain" this result states that some unknown factor causes the lake to acidify and also kills fish (i.e., the relationship between numbers of fish and acidification is due to a common response to another factor). However, experiments where lakes were deliberately acidified refute this alternate explanation. No such refutation is possible from surveys of existing populations. The primary message is that causation cannot be inferred without active manipulation.

Despite the weaker inferences from nonexperimental surveys, the same attention must be paid to the proper design of a survey so that inadvertent biases do not taint the conclusions. The many different types of nonexperimental surveys are outlined in Figure 3.1. To begin, consider the following series of examples to illustrate the differences among these types of surveys.

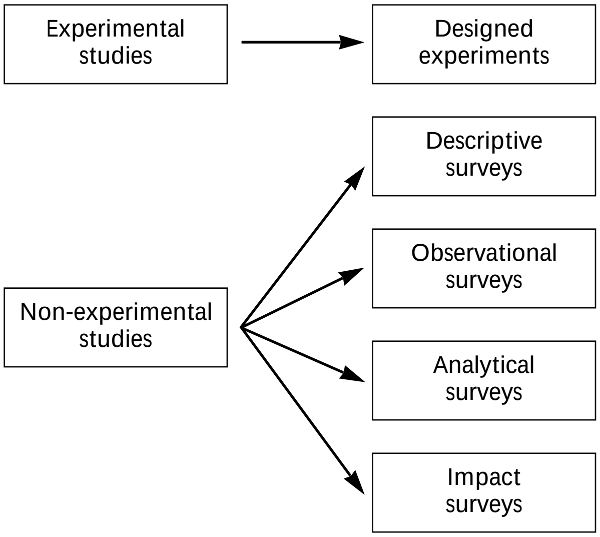


FIGURE 3.1 *A classification of the methods considered in this chapter.*

3.1.1 Example A: descriptive survey

A manager is interested in examining the natural regeneration in a cutblock harvested by clearcutting. The objective is to measure the amount of regeneration. A suitable response measure will be the density of newly grown trees. A series of sample plots is systematically located within a single cutblock and the density is measured on each sample plot. The mean density over all plots is computed along with a measure of precision, the standard error. The study has only one response variable, the density on each plot, and no explanatory variables. This is a *descriptive survey* as no comparisons will be made with other cutblocks and the information pertains only to that particular cutblock. No inferences about the density in other cutblocks is possible.

3.1.2 Example B: observational survey

This same manager now notices that north-facing slopes seem to have a lower insect infestation rates than south-facing slopes. One cutblock from a north-facing slope and one cutblock from a south-facing slope are selected. Sample plots are located on each cutblock, and the insect infestation is measured on each sample plot. The response variable is the amount of infestation in each plot. The orientation of the slope is an explanatory variable. Estimates of the mean infestation are obtained for each block. The sample means for each block likely differ, but with information on the variation within each block, it is possible to determine if the population means also differ (i.e., to determine if the true average infestation in the two blocks differs). This is an *observational survey* as two convenient cutblocks were selected and compared. However, the results are only applicable to the two cutblocks sampled and can neither be extrapolated to other cutblocks, nor to the effects of north- and south-facing slopes. The reason for this weak inference is that the observed differences between the cutblocks may be due to just natural variation unrelated to the direction of the slope; no information has been collected on the variability among cutblocks with the same orientation.

3.1.3 Example C: analytical survey

The manager expands the above survey. Within the forest management unit, blocks are randomly chosen in pairs so that, within each pair, one cutblock is on a north-facing slope and the other is on a south-facing slope. Sample plots are randomly located on each cutblock, and the insect infestation is measured on each sample plot. The response variable is the amount of infestation in each plot. The orientation is an explanatory variable. Estimates of the mean infestation are obtained for each type of slope along with the measures of precision. The manager then compares the two means using information on both the within-cutblock variability and the variability among blocks with the same orientation. It may appear that plots on the south-facing slope have a higher infestation than plots on a north-facing slope. This is an *analytical survey*, as a comparison was made over an entire population of cutblocks in the forest management unit. This survey differs from a controlled experiment in that the orientation of the cutblocks cannot be controlled by the manager. An alternative explanation for this observed result is that some other unknown factor caused the insect infestations to be different on the two orientations.

3.1.4 Example D: designed experiment

The manager is interested in testing the effect of two different types of fertilizer on regeneration growth. Experimental plots in several homogeneous cutblocks are established. Within each cutblock, plots are randomly assigned to one of the fertilizers. The regeneration growth of the plots treated with the two fertilizers is then compared. The response variable is the amount of growth; the explanatory variable is the fertilizer type. Because plots were randomly assigned to the fertilizers, the effects of any other, uncontrollable, lurking factor should, on average, be about equal in the two treatment groups. Consequently, any difference in the mean regeneration growth can be attributed to the fertilizer. The primary differences between this example and Example C are that the manager controls the explanatory factor and can randomly assign experimental units to treatments. These two differences in the protocol allow stronger inferences than in analytical surveys.

3.1.5 Example E: impact survey

The manager wishes to examine if clearcutting is changing the water quality on nearby streams. A control site with similar soil and topography as the experimental site, is selected, in a provincial park. Water quality readings are taken from both streams several times before harvesting, and several times after harvesting. The response variable is the water quality; the explanatory variable is the presence or absence of nearby clearcutting.

The changes in water quality in the control and experimental sites are compared. If the objective is to examine if there is a difference in water quality between these two specific sites, then the survey will answer the question. This is similar to the strength of inference for observational surveys (Example B). If the objective is to extrapolate from this pair of sites to the effects of clearcutting in general, the inference is much more limited. First, because the control or impact sites are not replicated it is impossible to know if the observed differences are within the range of natural variation. This limitation could be partly resolved by adding multiple control sites and assuming that the variability among control sites is representative of that among impact sites. However, the lack of randomization of the impact will still limit the extent to which the results can be generalized. But in the longer term, if there are several such pairs of sites and all show the same type of impact, solid grounds are established for assigning a causal relationship, even though randomization never took place. This would be based on the idea of a superpopulation consisting of all possible pairs of sites; it is not likely that unobservable, latent factors would be operating in the same direction in all experiments. This last form is the closest to a designed experiment for an impact survey.

These five examples differ in two important dimensions:

1. The amount of control over the explanatory factor. Descriptive surveys have the least amount of control, while designed experiments have maximal control.

2. The degree of extrapolation to other settings. Again, in descriptive surveys, inference is limited to those surveyed populations, while in designed experiments on randomly selected experimental units, inference can be made about future effects of the explanatory factors.

In general, the more control or manipulation present, the stronger the inferences that can be made (Figure 3.2).

This chapter will present an overview of some of the issues that arise in surveys lacking experimental manipulations. It will start with an overview of the descriptive surveys used to obtain basic information about a population. Observational surveys will not be explicitly addressed as their usefulness is limited and their design and analysis are very close to analytical surveys. Next, analytical surveys, where the goal is to compare subsets of an existing population, will be

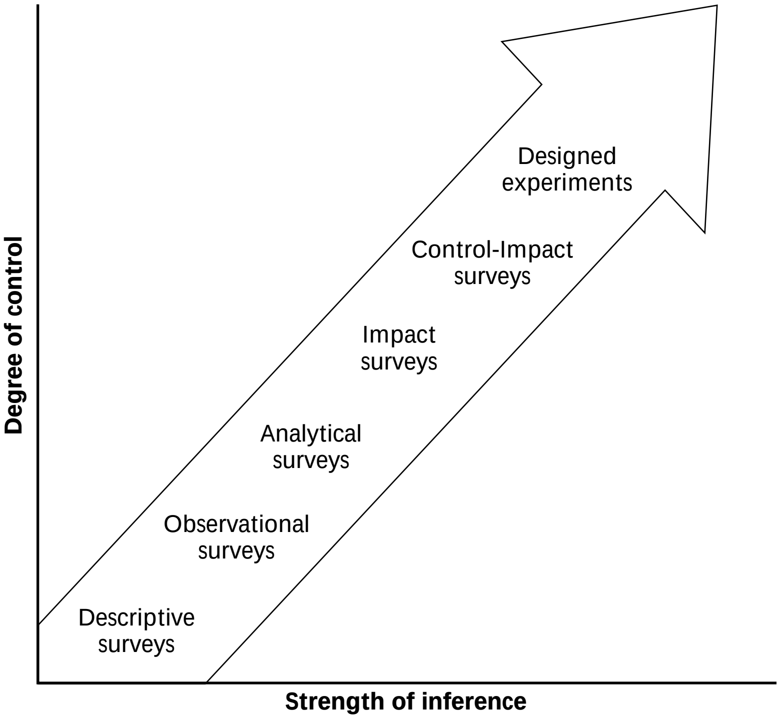


FIGURE 3.2 *Relationship between degree of control, strength of inference, and type of study design.*

described. Impact Surveys, where one site affected by some planned or unplanned event and a control site where no such event occurs are compared, will then be discussed. Finally, some general principles of non-experimental studies will be reviewed.

3.2 Descriptive Surveys

The goal of a descriptive survey is to estimate a parameter of interest (e.g., an average, total, proportion, or ratio) for a single population (e.g., Example A of Section 3.1). No attempt is made to compare the parameters between two or more populations.

Many excellent references on descriptive survey methods are available (Cochran 1977; Krebs, 1989; Thompson 1992). Therefore, this section is limited to a brief account of the main survey methods that could be used in field research. Details on actual field procedures are also available; for example, Myers and Shelton (1980).

3.2.1 Survey methods

Simple random sampling

Simple random sampling is the basic method of selecting survey units. Each unit in the population is selected with equal probability and all possible samples are equally likely to be chosen. This selection is commonly done by listing all the members in the population and then sequentially choosing units using a random number table. Units are usually chosen without replacement (i.e., each unit in the population can only be chosen once). In some cases, particularly for multistage designs, there are advantages to selecting units with replacement (i.e., a unit in the population may potentially be selected more than once).

The analysis of a simple random sample is straightforward. The mean of the sample is an estimate of the population mean. An estimate of the population total is obtained by multiplying the sample mean by the number of units in the population. The sampling fraction, the proportion of units chosen from the entire population, is typically small. If it exceeds 20%, an adjustment (the finite population correction) will result in better estimates of precision (a reduction in the standard error) to account for the fact that a substantial fraction of the population was surveyed.

An example of a simple random sample would be a vegetation survey in a large forest stand. The stand is divided into 300 1-hectare plots, and a random sample of 20 plots was selected and analyzed using aerial photos.

Pitfall: A simple random sample design is often "hidden" in the details of many other survey designs. For example, many surveys of vegetation are conducted using strip transects where the initial starting point of the transect is randomly chosen, and then every plot along the transect is measured. Here the strips are the sampling unit, and are a simple random sample from all possible strips. The individual plots are subsamples from each strip and cannot be regarded as independent samples. For example, suppose a rectangular stand is surveyed using aerial overflights. In many cases, random starting points along one edge are selected, and the aircraft then surveys the entire length of the stand starting at the chosen point. The strips are typically analyzed section-by-section, but it would be incorrect to treat the smaller parts as a simple random sample from the entire stand.

Solution: Note that a crucial element of simple random samples is that every sampling unit is chosen independently of every other sampling unit. For example, in strip transects, plots along the same transect are not chosen independently: when a particular transect is chosen, all plots along the transect are sampled and so the selected plots are not a simple random sample of all possible plots. Strip-transects are actually examples of cluster-samples.

Systematic sampling

In some cases, it is logistically inconvenient to randomly select sample units from the population. An alternative is to take a systematic sample where every k^{th} unit is selected (after a random starting point); k is chosen to give the required sample size. For example, if a stream is 2 km long, and 20 samples are required, then k=100 and samples are chosen every 100 m along the stream after a random starting point. A common alternative when the population does not naturally divide into discrete units is grid-sampling. Here sampling points are located using a grid that is randomly located in the area. All sampling points are a fixed distance apart.

If a known trend is present in the sample, it can be incorporated into the analysis (Cochran 1977, Chap. 8). For example, suppose that the systematic sample follows an elevation gradient that is known to directly influence the response variable. A regression-type correction can be incorporated into the analysis.

However, note that this trend must be known from external sources—it cannot be deduced from the survey.

Pitfall: A systematic sample is typically analyzed in the same fashion as a simple random sample. However, the true precision of an estimator from a systematic sample can be either worse or better than a simple random sample of the same size, depending if units within the systematic sample are positively or negatively correlated among themselves. For example, if a systematic sample's sampling interval happens to match a cyclic pattern in the population, values within the systematic sample are highly positively correlated (the sampled units may all hit the "peaks" of the cyclic trend), and the true sampling precision is worse than a simple random sample of the same size. What is even more unfortunate is that, because the units are positively correlated within the sample, the sample variance will underestimate the true variation in the population, and, if the estimated precision is computed using the formula for a simple random sample, a double dose of bias in the estimated precision occurs (Krebs 1989, p. 227). On the other hand, if the systematic sample is arranged "perpendicular" to a known trend to try to incorporate additional variability in the sample, the units within a sample are now negatively correlated, the true precision is now better than an simple random sample of the same size, but the sample variance now overestimates the population variance, and the formula for precision from a simple random sample will overstate the sampling error.

While logistically simpler, a systematic sample is only "equivalent" to a simple random sample of the same size if the population units are "in random order" to begin with (Krebs 1989, p. 227). Even worse, no information in the systematic sample allows the manager to check for hidden trends and cycles.

Nevertheless, systematic samples offer the following practical advantages over simple random sampling if the bias in the estimated precision can be corrected:

- make plot relocation for long-term monitoring easier;
- allow mapping to be carried out concurrently with the sampling effort because the ground is systematically traversed;
- avoid poorly distributed sampling units, which can occur with a simple random sample (this problem can also be avoided by judicious stratification).

Solution: Because a strong assumption of "randomness" in the original population is necessary, systematic samples are discouraged and statistical advice should be sought before starting such a scheme. If no other designs are feasible, a slight variation in the systematic sample provides some protection from the previous problems. Instead of taking a single systematic sample every k^{th} unit, take two or three independent systematic samples of every $2k^{\text{th}}$ or $3k^{\text{th}}$ unit, each with a different starting point. For example, rather than taking a single systematic sample every 100 m along the stream, two independent systematic samples can be taken, each selecting units every 200 m along the stream starting at two random starting points. The total sample effort is still the same, but now some measure of the large-scale spatial structure can be estimated. This technique is known as *replicated subsampling* (Kish 1965, p. 127).

Cluster sampling

In some cases, units in a population occur naturally in groups or clusters. For example, some animals congregate in herds or family units. It is often convenient to select a random sample of herds and then measure every animal in the herd. This is not the same as a simple random sample of animals because individual animals are not randomly selected; the herds were the sampling unit. The strip-transect example in Section 3.2.1 is also a cluster sample; all plots along a randomly selected transect are measured. The strips are the sampling units, while plots within each strip are subsampling units. Another example is circular plot sampling; all trees within a specified radius of a randomly selected point are measured. The sampling unit is the circular plot while trees within the plot are subsamples.

The reason cluster samples are used is that costs can be reduced compared to a simple random sample giving the same precision. Because units within a cluster are close together, travel costs among units are reduced. Consequently, more clusters (and more total units) can be surveyed for the same cost as a comparable simple random sample.

Pitfall: A cluster sample is often mistakenly analyzed using methods for simple random surveys. Such analysis is not valid because units within a cluster are typically positively correlated. This erroneous analysis produces an estimate that appears to be more precise than it really is (i.e., the estimated standard error is too small and does not fully reflect the actual imprecision in the estimate).

Solution: To be confident that the reported standard error really reflects the uncertainty of the estimate, the analytical methods must be appropriate for the survey design. The proper analysis treats the clusters as a random sample from the population of clusters. The methods of simple random samples are applied to the cluster summary statistics (Thompson 1992, Chap. 12; Nemec 1993).

Multi-stage sampling

In many situations the population is naturally divided into several different sizes of units. For example, a forest management unit consists of several stands, each stand has several cutblocks, and each cutblock can be divided into plots. These natural divisions can be easily accommodated in a survey through the use of multistage methods. Units are selected in stages. For example, several stands could be selected from a management area; then several cutblocks are selected in each of the chosen stands; then several plots are selected in each of the chosen cutblocks. Note that in a multistage design, units at any stage are selected at random only from those larger units selected in previous stages.

The advantage of multistage designs are that costs can be reduced compared to a simple random sample of the same size, primarily through improved logistics. The precision of the results is less than an equivalent simple random sample, but because costs are less, a larger multistage survey can often be completed for the same costs as a smaller simple random sample. This approach often results in a more precise design for the same cost. However, due to the misuse of data from complex designs, simple designs are often highly preferred and end up being more cost-efficient when costs associated with incorrect decisions are incorporated.

Pitfall: Although random selections are made at each stage, a common error is to analyze these types of surveys as if they arose from a simple random sample. The plots were not independently selected; if a particular cutblock was not chosen, then none of the plots within that cutblock can be chosen. As in cluster samples, this erroneous analysis produces estimated standard errors that are too small and do not fully reflect the actual imprecision in the estimates. A manager will be more confident in the estimate than is justified by the survey.

Solution: Again, it is important that the analytical methods are suitable for the sampling design. The proper analysis of multistage designs considers that random samples takes place at each stage (Thompson 1992, Chap. 13). In many cases, the precision of the estimates is determined essentially by the number of first-stage units selected. Little is gained by extensive sampling at lower stages.

Multiphase designs

In some surveys, multiple surveys of the same survey units are performed. In the first phase, a sample of units is selected (usually by a simple random sample). Every unit is measured on some variable. Then in subsequent phases, samples are selected ONLY from those units selected in the first phase, not from the entire population.

Multiphase designs are commonly used in two situations. In the first case, stratifying a population in advance is sometimes difficult because the values of the stratification variables are not known. The first phase is used to measure the stratification variable on a random sample of units. The selected units are then stratified, and each stratum is further sampled as needed to measure a second variable. This approach avoids having to measure the second variable on every unit when the strata differ in importance. For example, in the first phase, plots are selected and measured for the amount of insect damage. The plots are then stratified by the amount of damage, and second-phase allocation of units concentrates on plots with low insect damage to measure total usable volume of wood. It would be wasteful to measure the volume of wood on plots with heavy insect damage.

The second common occurrence for using a multistage design is a surrogate variable (related to the real variable of interest) on selected units is relatively easy to measure, and then, in the second phase, the real variable of interest is measured on a subset of the units. The relationship between the surrogate and desired variable in the smaller sample is used to adjust the estimate based on the surrogate variable in the larger sample. For example, managers need to estimate the volume of wood removed from a harvesting area. A large sample of logging trucks is weighed (which is easy to do), and weight will serve as a surrogate variable for volume. A smaller sample of trucks (selected from those weighed) is scaled for volume and the relationship between volume and weight from the second-phase sample is used to predict volume based on weight only for the first-phase sample. Another example is the count plot method of estimating volume of timber in a stand. A selection of plots is chosen and the basal area determined. Then a sub-selection of plots is rechosen in the second phase, and volumes are measured on the second-phase plots. The relationship between volume and area in the second phase is used to predict volume from area measurements seen in the first phase.

Repeated sampling

One common objective of long-term surveys is to investigate changes over time of a particular population. This investigation, which involves repeated sampling from the population, has three common designs.

First, separate independent surveys can be conducted at each time point. This is the simplest design to analyze because all observations are independent over time. For example, independent surveys can be conducted at 5-year intervals to assess regeneration of cutblocks. However, precision of the estimated change may be poor because of the additional variability introduced by having new units sampled at each time point.

At the other extreme, units are selected in the first survey and the same units are remeasured over time. For example, permanent plots that are remeasured for regeneration over time can be established. The advantage of permanent plots is that comparisons over time are free of additional variability introduced by new units being measured at every time point. One possible problem is that survey units have become "damaged" over time, and the sample size will tend to decline over time. An analysis of these types of designs is more complex because of the need to account for (1) the correlation over time of measurements on the same sample plot and (2) possible missing values when units become "damaged" and are dropped from the survey.

Intermediate to the previous two designs are partial replacement designs where a portion of the survey units is replaced with new units at each time point. For example, 20% of the units could be replaced by new units at each time point; units would normally stay in the survey for a maximum of five time periods. These types of designs require complex analysis.

Pitfall: The most common error made in analyzing repeated sampling designs is to treat observations as being independent. This typically leads to estimated precisions that appear too precise (i.e., the real precision is much poorer).

Solution: The analysis of repeated samples is quite complex—it is important to consult with an expert in this field.

Designs for wildlife sampling

Two common survey designs for measuring wildlife abundance are capture-recapture surveys and distance surveys.

In capture-recapture surveys (Otis et al. 1978; Pollock et al. 1990), animals are captured, tagged, and released on each of a number of time points. The pattern of recaptures of the observed animals is used to estimate survival rates and abundance. Skalski and Robson (1992) discuss the design of surveys using capture-recapture methods.

In distance surveys (Buckland et al. 1993), an observer follows a transect and notes the angle and distance of animals from the transect line. A detection function is constructed that relates the probability of spotting an animal as a function of the distance from the transect line and this is used to estimate abundance.

This section is deliberately brief as many complex planning problems are associated with using these methods and expert assistance is strongly recommended.

3.2.2 Refinements that affect precision

Sampling with unequal probability

All of the designs discussed in previous sections have assumed that each sample unit was selected with equal probability. In some cases, it is advantageous to select units with unequal probabilities, particularly if they differ in their contribution to the overall total. This technique can be used with any of the sampling designs discussed earlier. An unequal probability sampling design can lead to smaller standard errors (i.e., greater precision) for the same total effort compared to an equal probability design. For example, forest stands may be selected with probability proportional to the area of the stand (i.e., a stand of 200 ha will be selected with twice the probability than a stand of 100 ha) because large stands contribute more to the overall population and it would be wasteful to spend much sampling effort on smaller stands.

The variable used to assign the probabilities of selection to individual survey units does not need to have an exact relationship with an individual contribution to the total. For example, in probability proportional to prediction (3P) sampling, all trees in a small area are visited. A simple, cheap characteristic, which is used to predict the value of the tree, is measured. A subsample of the trees is then selected with probability proportional to the predicted value, remeasured using a more expensive measuring device. The relationship between the cheap and expensive measurement in the second phase is used with the simple measurement from the first phase to obtain a more precise estimate for the entire area. This example illustrates two-phase sampling with unequal probability of selection.

Stratification

All survey methods can potentially benefit from stratification (also known as blocking in the experimental-design literature). Stratification groups survey units into homogeneous groups before conducting the survey, and then conducts independent surveys in each stratum. At the end of the survey, the stratum results are combined and weighted appropriately. For example, a watershed might be stratified by elevation into three strata, and separate surveys are conducted within each elevation stratum. The separate results would be weighted proportionally to the size of the elevation strata. Stratification will be beneficial whenever variability among the sampling units can be anticipated and strata can be formed that are more homogeneous than the original population.

A major question with stratified surveys is the allocation of sampling units among the strata. Depending upon the goals of the survey, an optimal allocation of sampling units can be one that is equal in all strata, that is proportional to the stratum size, or that is related to the cost of sampling in each stratum (Thompson 1992, Chap. 11). Equal allocation (where all strata have the same sample size) is preferred when equally precise estimates are required for each stratum as well as for the overall population. Proportional allocation (where the sample size in each stratum is proportional to the population size) is preferred when more precise estimates are required in larger strata. If the costs of sampling vary among the strata, then an optimal allocation that accounts for costs would try to obtain the best overall precision at the lowest cost by allocating units among the strata accounting for the costs of sampling in each stratum.

Stratification can be carried out prior to the survey (pre-stratification) or after the survey (post-stratification). Pre-stratification is used if the stratum variable is known in advance for every plot (e.g., elevation of a plot). Post-stratification is used if the stratum variable can only be ascertained after measuring the plot (e.g., soil quality or soil pH). The advantages of pre-stratification are that samples can be allocated to the various strata in advance to optimize the survey and the analysis is relatively straightforward. With post-stratification, there is no control over sample size in each of the strata, and the analysis is more complicated (the samples sizes in each stratum are now random). Post-stratification can result in significant gains in precision but does not allow for finer control of the sample sizes as found in pre-stratification.

Auxiliary variables

An association between the measured variable of interest and a second variable of interest can be exploited to obtain more precise estimates. For example, suppose that growth in a sample plot is related to soil nitrogen content. A simple random sample of plots is selected and the height of trees in the sample plot is measured along with the soil nitrogen content in the plot. A regression model is fit (Thompson 1992, Chap. 7 and 8) between the two variables to account for some of the variation in tree height as a function of soil nitrogen content. This approach can be used to make precise predictions of the mean height in stands if the soil nitrogen content can be easily measured. This method will be successful if a direct relationship exists between the two variables. The stronger the relationship, the more effective this method will be. This technique is often called ratio-estimation or regression-estimation.

Notice that multiphase designs often use an auxiliary variable but this second variable is only measured on a subset of the sample units.

Unit size

A typical concern with any of the survey methods occurs when the population does not have natural discrete sampling units. For example, a large section of land may be arbitrarily divided into 1 m^2 plots, or 10 m^2 plots. A natural question—is what is the "best size" of unit?—has no simple answer and depends upon several factors, which must be addressed for each survey:

- Cost: All else being equal, sampling many small plots may be more expensive than sampling fewer larger plots. The primary difference in cost is the overhead in travel and setup to measure the unit.

- Size of unit: An intuitive feeling is that more smaller plots are better than few large plots because the sample size is larger. This will be true if the characteristic of interest is "patchy," but surprisingly, makes no difference if the characteristic is randomly scattered throughout the area (Krebs 1989, p. 64). Indeed if the characteristic shows "avoidance," then larger plots are better. For example, competition among trees implies they are spread out more than expected if they were randomly located. Logistical considerations often influence the plot size. For example, if trampling the soil affects the response, then sample plots must be small enough to measure without trampling the soil.
- Edge effects: Because the population does not have natural boundaries, decisions must often be made about objects that lie on the edge of the sample plot. In general, larger square or circular plots are better because of smaller edge-to-area ratio. (A long narrow rectangular plot can have more edge than a similar-area square plot.)
- Size of object being measured: Clearly, a 1 m^2 plot is not appropriate when counting mature Douglas-fir, but may be appropriate for a lichen survey.

A pilot survey should be carried out prior to a large-scale survey to investigate factors that influence the choice of sampling unit size.

Sample size determination

An important question in survey design is the choice of sample size, which is the primary determinant of the costs of the survey and of precision. The sample size should be chosen so that the final estimates have a precision that is adequate for the management question. Paradoxically, to determine the proper sample size, some estimate of the population values needs to be known before the survey is conducted! Historical data can sometimes be used. In some cases, pilot surveys will be needed to obtain preliminary estimates of the population values to plan the main survey. (Pilot surveys are also useful to test the protocol. Refer to Section 3.5).

Unfortunately, sometimes even pilot surveys cannot be done because of the difficulty in sampling or because the phenomenon is a one-time event. If a study has multiple objectives, reconciling the sample size requirements for each objective may also be difficult. In these and many other cases, sample sizes are determined solely by the budget for the survey.

3.3 Analytical Surveys

In descriptive surveys, the objective was to simply obtain information about one large group. In observational surveys, two deliberately selected subpopulations are chosen and surveyed, but the results are not generalized to the whole population. In analytical surveys, subpopulations are selected and sampled to generalize the observed differences among the subpopulation to this and other similar populations.

As such, analytical and observational surveys and experimental design are similar. However, the primary difference is that, in experiments, the manager controls the assignment of the explanatory variables while measuring the response variables, whereas in analytical and observational surveys, neither set of variables is under the control of the manager. (Refer to Section 3.1, Examples B, C, and D). The analysis of complex surveys for analytical purposes can be very difficult (Sedransk 1965a, 1965b, 1966; Rao 1973; Kish 1984, 1987).

The first step in analytical surveys is to identify potential explanatory variables (similar to factors in experiments). At this point, analytical surveys can be usually further subdivided into three categories depending on the type of stratification:

- the population is pre-stratified by the explanatory variables and surveys are conducted in each stratum to measure the outcome variables;
- the population is surveyed in its entirety, and post-stratified by the explanatory variables; and
- the explanatory variables can be used as auxiliary variables in ratio or regression methods.

In very complex surveys, all three types of stratification may take place.

The choice between the categories is usually made by the ease with which the population can be pre-stratified and the strength of the relationship between the response and explanatory variables. For example, sample plots can be easily pre-stratified by elevation or by exposure to the sun, but it would be difficult to pre-stratify by soil pH.

Pre-stratification has the advantage that the manager controls the number of sample points collected in each stratum. However, the numbers are not controllable in post-stratification and may lead to very small sample sizes in certain strata just because the strata form only a small fraction of the population.

For example, a manager may wish to investigate the difference in regeneration (as measured by the density of new growth) as a function of elevation. Several cutblocks will be surveyed. In each cutblock, the sample plots will be pre-stratified into three elevation classes, and a simple random sample will be taken in each elevation class. The allocation of effort in each stratum (i.e., the number of sample plots) will be equal. The density of new growth will be measured on each selected sample plot. On the other hand, suppose that the regeneration is a function of soil pH. This cannot be determined in advance, and so the manager must take a simple random sample over the entire stand, measure the density of new growth and the soil pH at each sampling unit, and then post-stratify the data based on measured pH. The number of sampling units in each pH class is not controllable—indeed it may turn out that certain pH classes have no observations.

If explanatory variables are treated as a auxiliary variables, then a strong relationship must exist between the response and explanatory variables and the auxiliary variable must be able to be measured precisely for each unit. Then, methods like multiple regression can also be used to investigate the relationship between the response and the explanatory variable. For example, rather than classifying elevation into three broad elevation classes or soil pH into broad pH classes, the actual elevation or soil pH must be measured precisely to serve as an auxiliary variable in a regression of regeneration density versus elevation or soil pH.

If the units have been selected using a simple random sample, then the analysis of the analytical surveys proceeds along similar lines as the analysis of designed experiments (Kish 1987; Nemec, this volume, Chap. 2). In most analyses of analytical surveys, the observed results are postulated to have been taken from a hypothetical super-population of which the current conditions are just one realization. In the above example, cutblocks would be treated as a random blocking factor, elevation class as an explanatory factor, and sample plots as samples within each block and elevation class. Hypothesis testing about the effect of elevation on mean density of regeneration occurs as if this were a planned experiment.

Pitfall: Any one of the sampling methods described in Section 3.2 for descriptive surveys can be used for analytical surveys. Many managers incorrectly use the results from a complex survey as if the data were collected using a simple random sample. As Kish (1987) and others have shown, this mistake can lead to substantial underestimates of the true standard error (i.e., the precision is thought to be far greater than is justified based on the survey results). Consequently, the manager may erroneously detect differences more often than expected (i.e., make a Type I error) and make decisions based on erroneous conclusions.

Solution: As in experimental design, it is important to match the analysis of the data with the survey design used to collect it. The major difficulties in analyzing analytical surveys are:

1. Recognizing and incorporating the sampling method used to collect the data in the analysis. The survey design used to obtain the sampling units must be taken into account in much the same way as the analysis of the collected data is influenced by actual experimental design. "Equivalencies" between terms in a sample survey and terms in experimental design are provided in Table 3.1. No quick and easy method is available for the analysis of complex surveys (Kish 1987). The super-population approach seems to work well if the selection probabilities of each unit are known (these are used to weight each observation appropriately) and if random effects corresponding to the various strata or stages are employed. The major difficulty caused by complex survey designs is that the observations are not independent of each other. This nonindependence, if properly incorporated into the analysis, can improve precision. If not accounted for, nonindependence will lead to seriously biased estimates of precision.

2. Unbalanced designs (e.g., unequal numbers of sample points in each combination of explanatory factors). This difficulty typically occurs if post-stratification is used to classify units by the explanatory variables but can also occur in pre-stratification if the manager decides not to allocate equal effort in each stratum. The analysis of unbalanced data is described by Milliken and Johnson (1984).

3. Missing cells (i.e., certain combinations of explanatory variables may not occur in the survey). The analysis of such surveys is complex, but refer to Milliken and Johnson (1984).

4. If the range of the explanatory variable is naturally limited in the population, then extrapolation outside of the observed range is not recommended.

More sophisticated techniques can also be used in analytical surveys. For example, correspondence analysis, ordination methods, factor analysis, multidimensional scaling, and cluster analysis all search for associations among measured variables that may give rise to hypotheses for further investigation. Unfortunately, most of these methods assume that units have been selected independently of each other using a simple random sample; extensions where units have been selected via a complex sampling design have not yet developed. Simpler designs are often highly preferred to avoid erroneous conclusions based on inappropriate analysis of data from complex designs.

Pitfall: While the analysis of analytical surveys and designed experiments are similar, the strength of the conclusions is not. In general, causation cannot be inferred without manipulation. An observed relationship in an analytical survey may be the result of a common response to a third, unobserved variable. For example, consider the two following experiments. In the first experiment, the explanatory variable is elevation (high or low). Ten stands are randomly selected at each elevation. The amount of growth is measured and it appears that stands at higher elevations have less growth. In the second experiment, the explanatory variable is the amount of fertilizer applied. Ten stands are randomly assigned to each of two doses of fertilizer. The amount of growth is measured and it appears that stands that receive a higher dose of fertilizer have greater growth. In the first experiment, the manager cannot say whether the differences in growth are due to differences in elevation or amount of sun exposure or soil quality as all three may be highly related. In the second experiment, all uncontrolled factors are present in both groups and their effects will, on average, be equal. Consequently, the assignment of cause to the fertilizer dose is justified because it is the only factor that differs (on average) among the groups.

As noted by Eberhardt and Thomas (1991), rigorous application of the techniques for survey sampling is needed when conducting analytical surveys, otherwise these surveys are likely to be subject to biases. Experience and judgement are very important in evaluating the prospects for bias, and attempting to find ways to control and account for these biases. The most common source of bias is the selection of survey units; the most common pitfall is to select units based on convenience rather than on a probabilistic sampling design. The potential problems that this can lead to are analogous to those that occur when it is assumed that callers to a radio phone-in show are representative of the entire population.

TABLE 3.1 *Equivalencies between terms used in surveys and in experimental design*

Survey term	Experimental design term
Simple random sample	Completely randomized design
Cluster sampling	(a) Clusters are random effects; units within a cluster treated as subsamples; or (b) Clusters treated as main plots; units within a cluster treated as subplots in a split-plot analysis
Multi-stage sampling	(a) Nested designs with units at each stage nested in units in higher stages. Effects of units at each stage treated as random effects; or (b) Split-plot designs with factors operating at higher stages treated as main plot factors and factors operating at lower stages treated as subplot factors
Stratification	Fixed factor or random block depending on the reasons for stratification
Sampling unit	Experimental unit or treatment unit
Subsample	Subsample

3.4 Impact Surveys

Probably the most important and controversial use of surveys is to investigate the effects of large-scale, potentially unreplicated events. These impact surveys investigate the impact of an event or process. In many cases, this survey must be done without having the ability or resources to conduct a planned experiment.

Consider three examples: the impact of a hydroelectric dam on water quality of the dammed stream; the impact of clearcuts on water quality of nearby streams; and the effect of different riparian zone widths along streams near clearcuts. First, randomization and replication are not possible in the first example. Only one dam will be built on one stream. In the other two examples, it is possible to randomize and replicate the experiment and so the principles of experimental design may be useful. Second, the impact of the first two examples can be compared to a control or non-treated site while in the last example impacts are compared: the two different riparian zone widths.

Regardless of the control over randomization and replication, the goal of impact surveys is typically to measure ecological characteristics (usually over time) to look for evidence of a difference (impact) between the two sites. Presumably, this impact will be attributed to the event, but, as shown later, the lack of replication and randomization may limit the generalization of the findings. Then, based on the findings, remediation or changes in future events will be planned. In all cases, the timing of the event must be known in advance so that baseline information can be collected.

A unifying example for this section will be an investigation of the potential effects of clearcuts on water quality of nearby streams. Several, successively more complex impact designs will be considered.

3.4.1 Designs

Before-after contrasts at a single site

This is the simplest impact design. A single survey is taken before and after a potential disturbance. This design is widely used in response to obvious accidental incidences of potential impact (e.g., oil spills, forest fires), where, fortuitously, some prior information is available. From this survey, the manager obtains a single measurement of water quality before and after the event. If the second survey reveals a change, this difference is attributed to the event.

Pitfalls: The observed event and the changes in the response variable may not be related—the change may be entirely coincidental. Even worse, no information is collected on the natural variability of the water quality over time and the observed differences may simply be due to natural fluctuations over time. Decisions based on this design are extremely hard to justify. This design cannot be used if the event cannot be planned and no prior data are available. In these cases, little can be said about the impact of the event.

Repeated before-after sampling at a single site

An embellishment on the previous sampling scheme is to perform multiple surveys of the stream at multiple time points before and after the event. In this design, information is collected on the mean water quality before and after the impact. As well, information is collected on the natural variability over time. This design is better than the previous design in that observed changes due solely to natural fluctuations over time can be ruled out and consequently any observed change in the mean level is presumably real.

The choice between regular intervals and random intervals depends upon the objectives of the survey. If the objective is to detect changes in trend, regularly spaced intervals are preferred because the analysis is easier. On the other hand, if the objective is to assess differences before and after impact, then samples at random time points are advantageous, because no cyclic differences unforeseen by the sampler will influence the size of the difference. For example, surveys taken every summer for several years before and after the clearcutting may show little difference in water quality but potentially significant differences in the winter may go undetected.

Pitfall: Despite repeated surveys, this design suffers from the same flaw as the previous design. The repeated surveys are pseudoreplications in time rather than real replicates (Hurlbert 1984). The observed change may have occurred regardless of the clearcut because of long-term trends over time. Again, decisions based on this design are difficult to justify.

BACI: Before-after-control-impact surveys

As Green (1979) pointed out, an optimal impact survey has several features:

- the type of impact, time of impact, and place of occurrence should be known in advance;
- the impact should not have occurred yet; and
- control areas should be available.

The first feature allows the surveys to be efficiently planned to account for the probable change in the environment. The second feature allows a baseline survey to be established and to be extended as needed. The last feature allows the surveyor to distinguish between temporal effects unrelated to the impact and changes related to the impact.

The simplest BACI design will have two times of sampling (before and after impact) in areas (treatment and a control) with biological and environmental variables being measured in all combinations of time and area. In this example, two streams would be sampled. One stream would be adjacent to the clearcut (the treatment stream); the second stream would be adjacent to a control site that is not clearcut and should have similar characteristics to the treatment stream and be exposed to similar climate and weather. Both streams are sampled at the same time points before the clearcut occurs and at the same time point after the clearcut takes place. Technically, this is known as an area-by-time factorial design, and evidence of an impact is found by comparing the before and after samples for the control site with the before and after samples for the treatment sites. This contrast is known as the area-by-time interaction (see Figure 3.3).

This design allows for both natural stream-to-stream variation and coincidental time effects. If the clearcut has no effect, then change in water quality between the two time points should be the same (i.e., parallel lines in Figures 3.3a and b). On the other hand, if the clearcut has an impact, the time trends will not be parallel (Figures 3.3c, d, and e).

Pitfalls: Hurlbert (1984), Stewart-Oaten et al. (1986), and Underwood (1991) discuss the simple BACI design and point out concerns with its application.

First, because impact to the sites was not randomly assigned, any observed difference between control and impact sites may be related solely to some other factor that differs between the two sites. One could argue that it is unfair to ascribe the effect to the impact. However, as Stewart-Oaten et al. (1986) point out, the survey is concerned about a particular impact

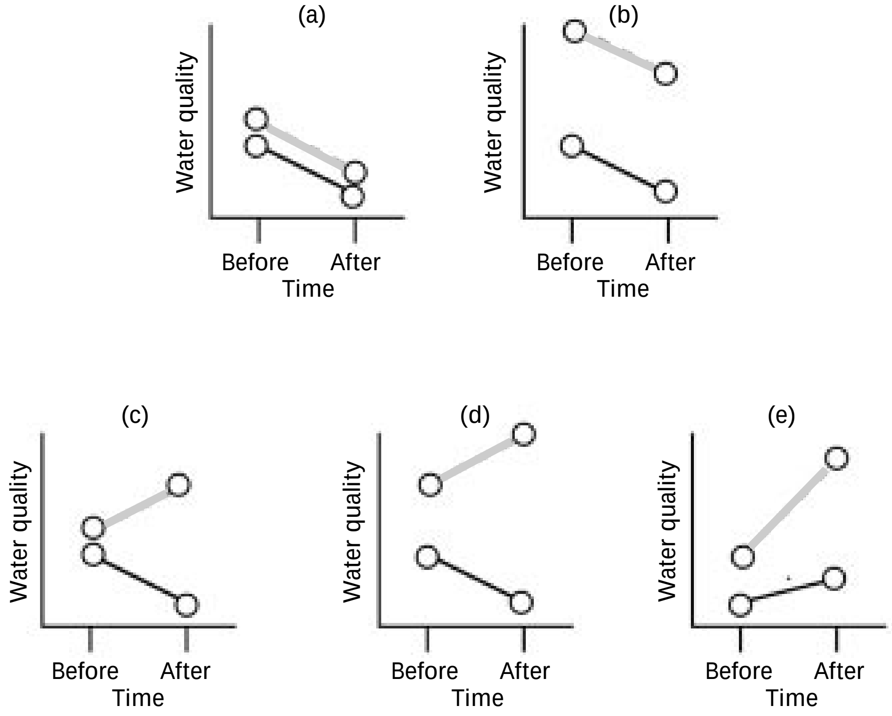


FIGURE 3.3 *Simplified outcomes in a BACI design.*
The change in a measured variable from two sampling occasions (dots at before and after the impact) in the control (solid line) or impact (shaded line) sites. In (a) and (b) the lines are parallel and there is no evidence of an impact. The difference in (b) between control and impact sites reflects area differences, but both sites experience the same temporal trend. In (c), (d), and (e), the change over time differs between the control and impact sites. This change is evidence of a time-treatment interaction, or that the impact has had an effect.

in a particular place, not in the average of the impact when replicated in many different locations. Consequently, detecting a difference between these two specific sites may be possible; however, without randomization of replicate treatments at many different sites, the findings from this survey cannot be generalized to other events on different streams.

This concern can be reduced by monitoring several control sites (Underwood 1991). However, two assumptions must be made: (1) the variation in the (After−Before) measurements of the multiple control sites is the same as the variation among potential impact sites, and (2) the variability over time between the control sites is not correlated. Then the plausability of the difference observed in the impact site can be estimated given the observed variability in the changes in the control sites. In our example, several control streams could be monitored at the same time points as the single-impact stream. Then if the observed difference in the impact stream is much different than could be expected based on the multiple-control streams, the event is said to have caused an impact. When several control sites are monitored, the lack of randomization is less of a concern because the replicated control sites provide some information about potential effects of other factors.

The second and more serious concern with the simple BACI design with a single sampling point before and after the impact is that it fails to recognize that natural fluctuations in the characteristic of interest that are unrelated to any impact may occur (Hurlbert 1984; Stewart-Oaten et al. 1986). For example, consider Figure 3.4. If there were no natural fluctuations over time, the single samples before and after the impact would be sufficient to detect the effects of the impact. However, if the population also

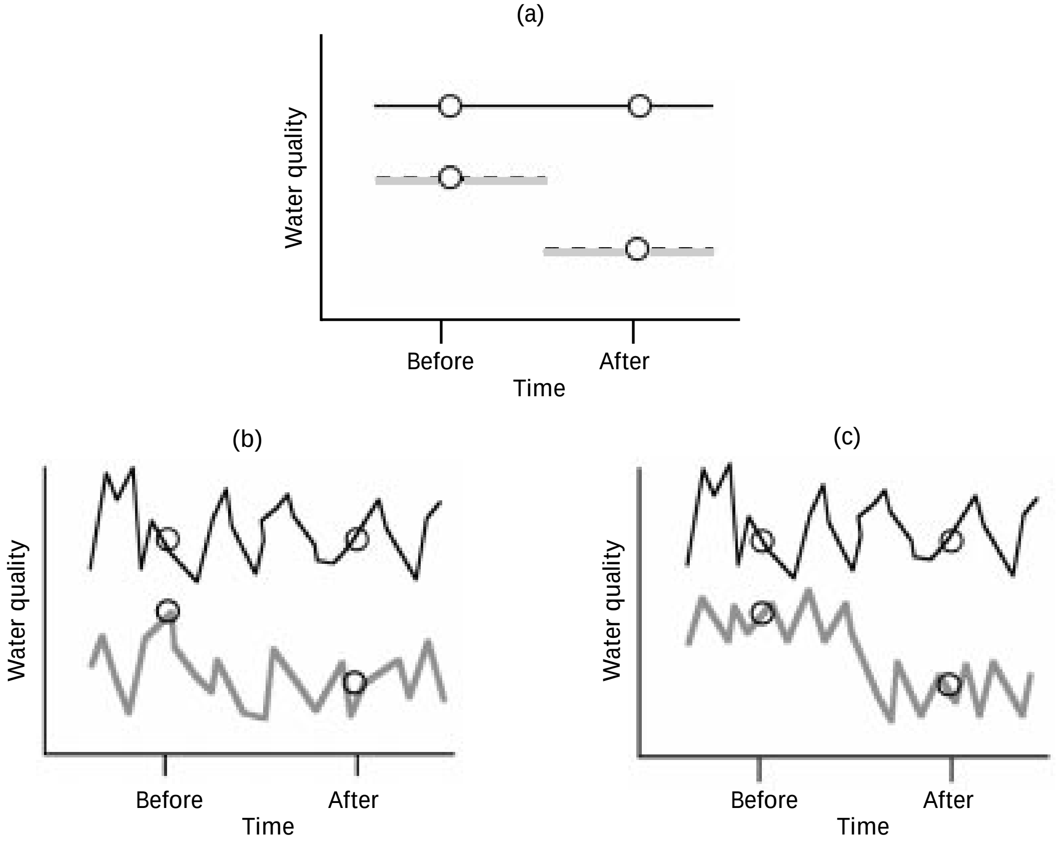


FIGURE 3.4 *Problems with the simple BACI design.*
The change in a measured variable from two sampling occasions (dots at before and after the impact) in the control (solid line) or impact (shaded line) sites. In (a), there is little natural variation in the response over time and so the measured values indicate a change in the mean level. In (b) and (c), natural variation is present, but, because only one point was sampled before and after impact, it is impossible to distinguish between no impact (b) and impact (c) on the mean level.

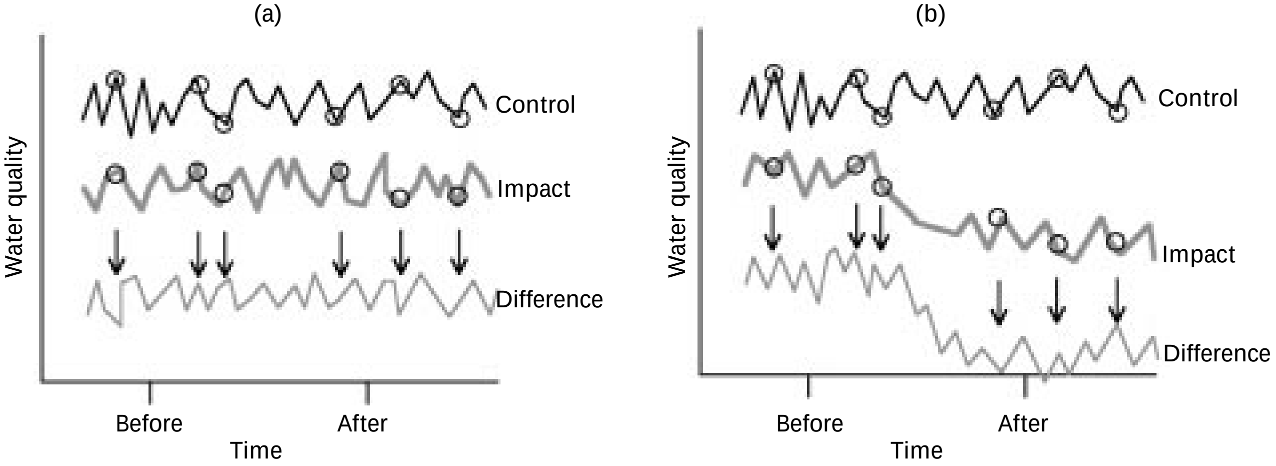


FIGURE 3.5 *The BACI-P design.*
The change in a measured variable from multiple randomly chosen sampling occasions (dots at before and after the impact) in the control (solid line) or impact (shaded line) sites. In (a), there is no impact and the mean level of the difference (bottommost line) is constant over time. In (b), there is an impact, and the mean level of the difference (bottommost line) changes over time.

has natural fluctuations over and above the long-term average, then distinguishing between cases where there is no effect from those where there was impact is impossible. In terms of our example, differences in the water quality may be artifacts of the sampling dates and natural fluctuations may obscure differences or lead to the conclusion that differences are present when they are not.

BACI-P: Before-after-control-impact paired designs

Stewart-Oaten et al. (1986) extended the simple BACI design by pairing surveys at several selected time points before and after the impact. Both sites are measured at the same time points. An analysis of how the difference between the control and impact sites changes over time would reveal if an impact has occurred (Figure 3.5). The rationale behind the design is that repeated sampling before the development indicates the pattern of differences over several periods of potential change between the two sites. This survey design provides information both on the *mean difference* in the water quality before and after impact, and on the *natural variability* of the water quality measurements. If the changes in the mean difference are large relative to natural variability, the manager has detected an effect.

The decision between random and regularly spaced intervals has been discussed in an earlier section—the same considerations apply here.

Pitfall: As with all surveys, numerous assumptions need to be made during the analysis (Stewart-Oaten et al. 1992; Smith et al. 1993). The primary assumption is that the responses over time are independent of each other. A lack of independence over time tends to produce false-positives (Type I errors) where the manager may declare that an impact has occurred when in fact, none has. In these cases formal time series methods may be necessary (Rasmussen et al. 1993). (The analysis of time series is easiest with regularly spaced sampling points).

Two other assumptions are made: that the difference in mean level between control and impact sites is constant over time in the absence of an impact effect and that the effect of the impact is to change the arithmetic difference. In our example, the difference in the mean water quality between the two sites would be assumed to be constant over time. The mean water quality measurements may fluctuate over time, but both sites are assumed to fluctuate in lockstep with each other maintaining the same average arithmetic difference. One common way this assumption is violated is if the response variable at the control site is a constant multiple of the response variable at the impact site. Then arithmetic differences will depend upon the actual levels. For example, suppose that the readings of water quality at two sites at the first time point were 200 versus 100, which has an arithmetic difference of 100; at the second time point, the readings were 20 versus 10, which has an arithmetic difference of 10; but both pairs are in a 2:1 ratio at both time points. The remedy is simple: a logarithmic transform of the raw data converts a multiplicative difference into a constant arithmetic difference on the logarithmic scale. This problem is commonly found when water quality measurements are concentrations (e.g., pH).

Underwood (1991) also considered two variations on the BACI-P design. First, it may not be possible to sample both sites simultaneously for technical or logistical reasons. Underwood (1991) discussed a modification where sampling is done at different times in each site before and after impact (i.e., sampling times are no longer paired), but notes that this modification cannot detect changes in the two sites that occurred before the impact. For example, differences in water quality may show a gradual change over time in the paired design prior to impact. Without paired sampling, it would be difficult to detect this change. Second, sampling only a single control site still has the problems identified earlier of not knowing if observed differences in the impact and the control sites are site-specific. Again, Underwood (1991) suggests that multiple control sites should be monitored. In our example, more than one control site would be measured at each time point. The variability in the difference between each control site and the impact site provides information on generalization to other sites.

Enhanced BACI-P: Designs to detect acute versus chronic effects or to detect changes in variation as well as changes in the mean

As Underwood (1991) pointed out, the previous designs are suitable for detecting long-term (chronic) effects in the mean level of some variable. In some cases, the impact may have an acute effect (i.e., effects only last for a short while) or may change the variability in response (e.g., seasonal changes become more pronounced). Underwood's solution is to modify the sampling schedule so that it occurs on two temporal scales (Figure 3.6). For example, groups of surveys could be conducted every 6 months with three surveys 1 week apart randomly located within each group. The analysis of such a design is presented in Underwood (1991). Again, several control sites should be used to confound the argument about detected differences being site-specific.

This design is also useful when there are different objectives. For example, the objective for one variable may be to detect a change in trend. The pairing of sample points on the long time scale leads to efficient detection of trend changes. The objectives for another variable may be to detect differences in the mean level. The short time scale surveys randomly located

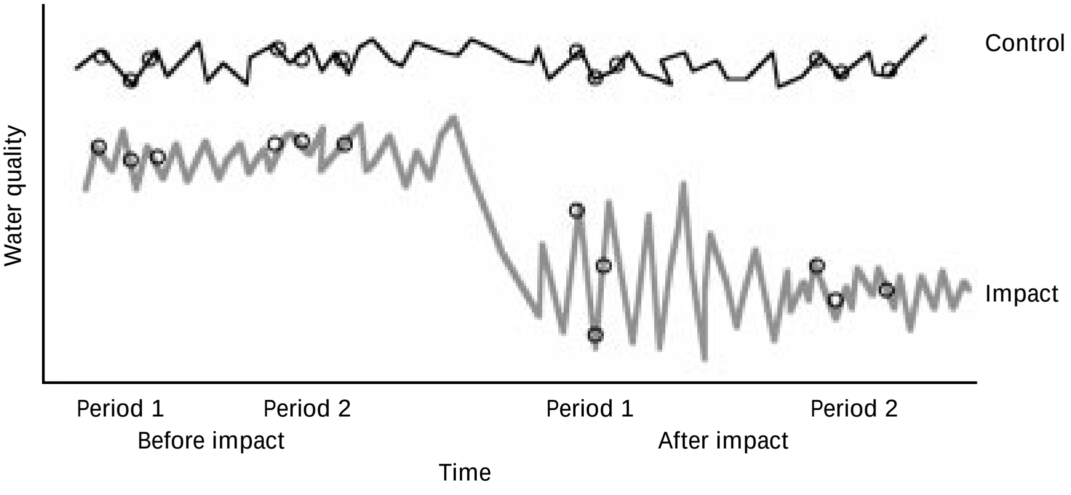


FIGURE 3.6 *The enhanced BACI-P design.*
The change in a measured variable from multiple randomly chosen sampling occasions in two periods (dots at before and after the impact) in the control (black line) or impact (shaded line) sites. The two temporal scales (sampling periods vs sampling occasions) allows the detection of a change in mean and in a change in variability after impact.

in time and space are efficient for detecting differences in the mean level.

3.4.2 Issues in impact surveys

Time dependence

Many of the analyses proposed for the above surveys (e.g., regression or ANOVA) have methodological problems that need to be resolved before interpreting the results.

In regression of the characteristics versus time, the estimated slope is often used as evidence of a long-term change. However, data collected over time violate the assumption of independence required for ordinary regression. The estimate of the slope remains unbiased, but typically the estimated standard error of the slope is too small. The results appear to be "statistically significant" when, in fact, there is no evidence of a change (Neter et al. 1990, Chap. 13) and a Type I error would have been made.

Comparing the means before and after impact using ANOVA methods also suffers from the same problem of correlation among the measurements. Again, a typical result is that the estimated standard error of the difference is too small, and results are declared "statistically significant" when in fact they are not, and a Type I error would have been made.

An alternative analysis is to use time-series methods that incorporate temporal correlation. The analysis of time series is quite complex (Nelson 1973) particularly if the time points are unequally spaced. If the data points are taken before and after the impact, the time series analysis can be extended using intervention analysis to test if an impact changed the level of the series (Rasmussen et al. 1993).

Temporary or permanent monitoring sites

A common question in monitoring surveys is the use of temporary or permanent monitoring sites. For example, should permanent water quality sampling sites that are remeasured over time, or temporary sampling sites that are re-randomized at each time be used? Many of the concerns are similar to those for repeated sampling designs discussed earlier. Permanent plots give better estimates of change over time because the extra plot-to-plot variability caused by bringing in new plots each year is removed. However, the costs of establishing permanent plots are higher than for temporary sites, and the first randomization may lead to a selection of plots that have some strange characteristics. Of course, if the measurement process alters the sampling unit, new plots will have to be selected for each survey. A compromise solution is a rotating panel survey, where only a part of the sample is changed at each time point. In large, complex, long-term designs with multiple objectives, permanent plots are often the preferred solution since no survey design is optimal for all objectives and the objectives change over time.

3.4.3 Impact surveys summary

As noted by Smith et al. (1993), the BACI-P design and its extensions are one of the best models for impact assessment. These designs can show that observed differences in ecological variables between the control and impact sites are neither artifacts of sampling nor due to temporal trends unrelated to the impact. The strength of the inference is directly related to the design issues directly under the control of the managers such as the frequency of sampling and number of control sites. Because of the potentially large amounts of data collected, quality assurance methods need to be employed throughout the length of the survey so that problems in data management, data handling, or changes in personnel do not compromise the survey.

3.5 Conclusion

Green (1979) gave 10 principles applicable to any sampling design; these principles have been paraphrased, reordered, and extended below. Underwood (1994) also gives some advice on areas of common misunderstanding between environmental biologists and statisticians.

1. Formulate a clear, concise hypothesis

The success or failure of a sampling program often hinges on clear, explicit hypotheses. Woolly thinking at this stage frequently leads to massive amounts of data collected without enough planning as to how, to what end, and at what cost the information can be subsequently handled. Hypotheses should be stated in terms of direct, measurable variables (e.g., action X will cause a decrease in Y). The hypotheses to be tested have implications for what and how data are to be collected.

2. Ensure that controls will be present

Most surveys are concerned with changes over time, typically before and after some impact. Effects of an impact cannot be demonstrated without the presence of controls serving as a baseline so that changes over time, unrelated to the impact, can be observed. Without controls, no empirical data are available to refute the argument that observed changes might have occurred regardless of impact.

3. Stratify in time and space to reduce heterogeneity

If the area to be sampled is large and heterogeneous (highly variable), then sampling from the entire area, ignoring the known heterogeneity, reduces the precision of the estimate. Extra variation may be introduced to the measured variable solely by differences within the survey area unrelated to the treatment. By stratifying the survey area in advance (also known as blocking in the experimental design literature), this extra variability can be accounted for. The judicious choice of auxiliary variables can also be used to increase precision of the estimates.

4. Take replicate samples within each combination of time, space, or any other controlled variable

Differences among treatments can only be demonstrated by comparing the observed differences among treatments with differences within each treatment. Lack of replication often restricts the interpretation of many experiments and surveys to the sampled units rather than to the entire population of interest. It is imperative that the replicates be true replicates and not pseudoreplicates (Hurlbert 1984), where the same experimental unit is often measured many times.

5. Determine the size of a biologically meaningful, substantive difference that is of interest

A sufficiently large survey (i.e., with large sample sizes) can detect minute differences that may not be of biological interest. It is important to quantify the size of a difference that is biologically meaningful before a survey begins so that resources are not wasted either by performing a survey with an excessive sample size or by performing a survey that has lower power to detecting this important difference.

6. Estimate the required sample sizes to obtain adequate power to detect substantive differences or to ensure sufficient precision of the estimates

In this era of fiscal restraint, it is unwise to spend significant sums of money on surveys or experiments that have only a slight chance of detecting the effect of interest or give estimates that are so imprecise as to be useless. Such designs are a waste of time and money.

If the goal of the survey is to detect a difference among populations, the required sample sizes will depend upon the magnitude of the suspected difference, and the amount of natural variation present. Estimates of these qualities can often be obtained from experience, literature reviews of similar surveys, or pilot surveys. Simulation studies can play an important role in assessing the efficiency of a design.

If the goal is descriptive, then the required sample sizes will depend only upon the natural variation present. As mentioned, estimates of the variability can be obtained from experience, literature reviews, or a pilot survey.

As noted earlier, it may be infeasible to conduct a pilot survey, historical data may not exist, or it may be difficult to reconcile sample sizes required for different objectives. Some compromise will be needed (Cochran 1977, pp. 81–82).

One common misconception is that sample size is linked to the size of the population. To the contrary, the sample sizes required to estimate a parameter in a small population with a specified precision are the same as in a large population. This non-intuitive result has a direct analogue in testing a pot of soup for salt—the cook tastes only a spoonful regardless of pot size.

7. Allocate replicate samples using probabilistic methods in time and space

There is a tendency to allocate samples into "representative" or "typical" locations. Even worse are convenience samples where the data are collected at sampling points that are easily accessible or close-at-hand. The key to ensuring "representativeness" is randomization. Randomization ensures that the effects of all other uncontrollable variables are equal, on average, in the various treatment groups or that they appear in the sample, on average, in the same proportions as in the population. Unless the manager is omniscient, it is difficult to ensure that "representative" or "typical" sites are not affected by other, unforeseen, uncontrollable factors.

Notice that a large sample size does not imply representativeness. Randomization controls representativeness; sample size controls statistical power.

8. Pretest the sampling design and sampling methods

It is difficult to spend effort on a pilot survey knowing that the data collected may not contribute to the final results and may be thrown away. However, this approach is the only way to check if serious problems exist in the survey, if the size of the survey unit is appropriate, if the data collection forms are adequate, and if the actual level of variability is present in the field, etc.

After a pilot survey has been conducted, its results can be used to modify the proposed design and fine-tune such aspects as the required sample size. In many cases, a pilot survey shows that the objectives of the proposed survey are unobtainable for the projected cost and effort and the survey must be substantially modified or abandoned.

9. Maintain quality assurance throughout the survey
Despite best efforts, plans will deviate during the course of the survey, particularly if the survey extends over many years and personnel changes. Many of the principles of statistical process control can be applied here (Montgomery 1991). For example, ensure that instruments are recalibrated at regular intervals, sampling protocols are followed consistently among different team members, and data are being keyed correctly.

10. Check the assumptions of any statistical analysis
Any statistical procedure makes explicit and implicit assumptions about the data collected. Match the analysis with the survey design. In many cases, a "statistically significant" result can be obtained erroneously if assumptions necessary for the analysis were violated.

11. Use the "Inter-Ocular Trauma Test"
Presentation of final results is just as important as design, execution, and analysis. A survey will be of limited usefulness if it sits on a shelf because other readers are unable to interpret the findings. Good graphical methods (figures, plots, charts, etc.) or presentations will pass the Inter-Ocular Trauma Test (i.e., the results will "hit you between the eyes!")

Despite their limitations, uncontrolled events can play a useful role in adaptive management. The study of uncontrolled events and designed experiments differ in two important dimensions:

1. The amount of control. As the name implies, the study of uncontrolled events does not give the manager the ability to manipulate the explanatory variables.
2. The degree of extrapolation to other settings. The lack of randomization implies that the manager must be careful in extrapolating to new situations because of the possible presence of latent, lurking factors.

These differences imply that inferences are not as strong as those made after carefully controlled experiments, but the results often lead to new hypotheses being tested in future research. Despite the weaker inferences from studying uncontrolled events, the same attention must be paid to the proper design of a survey so that inadvertent biases do not taint the conclusions.

References

Box, G.E.P. and G.C. Tiao. 1975. Intervention analysis with applications to economic and environmental problems. J. Am. Statist. Assoc. 70:70–9.

Buckland, S.T., D.R. Anderson, K.P. Burnham, and J.L. Laake. 1993. Distance sampling: estimating abundances of biological populations. Chapman and Hall, London, U.K.
The standard monograph on the statistical analysis of distance sampling experiments.

Cochran, W.G. 1977. Sampling techniques. J. Wiley, New York, N.Y.
One of the standard references for survey sampling. Very technical.

Eberhardt, L.L. and J.M. Thomas 1991. Designing environmental field studies. Ecol. Monogr. 61:53–73.
An overview of the eight different field situations as shown in Figure 1.

Fletcher, D.J. and B.F.J. Manly (editors). 1994. Statistics in ecology and environmental monitoring. Univ. Otago Press, Dunedin, N.Z.
A collection of papers at a moderate to advanced level on a wide range of topics.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. J. Wiley, New York, N.Y.
One of the first comprehensive unified treatments of sampling issues for environmental biologists. Very readable.

______. 1993. Application of repeated measures designs in environmental impact and monitoring studies. Austr. J. Ecol. 18:81–98.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 52:187–211.
A critique of many common problems encountered in ecological field experiments.

Keith, L.H. (editor). 1988. Principles of environmental sampling. Am. Chem. Soc., New York, N.Y
A series of papers on sampling mainly for environmental contaminants in ground and surface water, soils, and air. A detailed discussion on sampling for pattern.

Kish, L. 1965. Survey sampling. J. Wiley, New York, N.Y.
An extensive discussion of descriptive surveys mostly from a social science perspective.

______. 1984. On analytical statistics from complex samples. Sur. Methodol. 10:1–7.
An overview of the problems in using complex surveys.

______. 1987. Statistical designs for research. J. Wiley, New York, N.Y.
One of the more extensive discussions of the use of complex surveys in analytical surveys. Very technical.

Krebs, C.J. 1989. Ecological methodology. Harper and Row, New York, N.Y.
A methods books for common techniques used in ecology.

Milliken, G.A. and D.E. Johnson 1984. The analysis of messy data: Vol. 1: designed experiments. Van Nostrand Reinhold, New York, N.Y.
A complete treatise on the analysis of unbalanced data in designed experiments. Requires a background in the use of ANOVA methodology.

Montgomery, D.C. 1991. Introduction to statistical quality control. J. Wiley, New York, N.Y.
A standard introduction to the principles of process control.

Myers, W.L. and R.L. Shelton 1980. Survey methods for ecosystem management. J. Wiley, New York, N.Y.
Good primer on how to measure common ecological data using direct survey methods, aerial photography, etc. Includes a discussion of common survey designs for vegetation, hydrology, soils, geology, and human influences.

Nelson, C.R. 1973. Applied time series analysis for managerial forecasting. Holden-Day, San Francisco, Calif.
A primer on the basic time series analysis methods.

Nemec, A.F.L. 1993. Standard error formulae for cluster sampling (unequal cluster sizes). B.C. Min. For., Res. Br., Victoria, Biometric Inf. Pamph. No. 43.

______. [n.d.]. Design of experiments. This volume.

Neter, J.N., W. Wasserman, and M.H. Kutner. 1990. Applied linear statistical models: regression, analysis of variance, and experimental designs, 3rd ed. Irwin, Boston, Mass.
A standard treatment of regression and experimental design suitable after a first course in statistics.

Otis, D.L., K.P. Burnham, G.C. White, and D.R. Anderson. 1978. Statistical inference from capture–data on closed animal populations. Wildl. Monogr. 62.
The standard monograph on the statistical analysis of mark-recapture experiments in closed populations.

Pollock, K.H., J.D. Nichols, C. Brownie, and J.E. Hines 1990. Statistical inference from capture–recapture experiments. Wildl. Monogr. 107.
The standard monograph on the statistical analysis of mark-recapture experiments in open populations.

Rao, J.N.K. 1973. On double sampling for stratification and analytical surveys. Biometrika 60:125–33.

Rasmussen, P.W., D.M. Heisey, E.V. Nordheim, and T.M. Frost 1993. Time series intervention analysis: unreplicated large-scale experiments. *In* Design and analysis of ecological experiments. S.M. Scheiner and J. Gurevitch (editors). pp. 138–58. Chapman and Hall, New York, N.Y.

Scheiner, S.M. and J. Gurevitch 1993 (editors). Design and analysis of ecological experiments. Chapman and Hall, New York, N.Y.

Sedransk, J. 1965a. A double sampling scheme for analytical surveys. J. Am. Statist. Assoc. 60:985–1004.

______. 1965b. Analytical surveys with cluster sampling. J. Royal Statist. Soc., B, 27:264–78.

______. 1966. An application of sequential sampling to analytical surveys. Biometrika 53:85–97.

Skalski, J.R. and D.S. Robson 1992. Techniques for wildlife investigations: design and analysis of capture data. Academic Press, New York, N.Y.
Presents methods for conducting experimental inference and mark-recapture statistical studies for fish and wildlife investigations.

Smith, E.P., D.R. Oruos, and J. Cairns Jr. 1993. Impact assessment using the before-after-control-impact (BACI) model: concerns and comments. Can. J. Fisheries Aquatic Sci. 50:627–37

Stewart-Oaten, A., J.R. Bence, and C.W. Osenberg. 1992. Assessing effects of unreplicated perturbations—no simple solutions. Ecology 73:1396–404.

Stewart-Oaten, A., W.M. Murdoch, and K. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67:929–40.
One of the first extensions of the BACI design discussed in Green (1979).

Thompson, S.K. 1992. Sampling. J. Wiley, New York, N.Y.
A good companion to Cochran (1977). Has many examples of using sampling for biological populations. Also has chapters on mark-recapture, line-transect methods, spatial methods, and adaptive sampling.

Underwood, A.J. 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Austr. Marine and Freshwater Res. 42:569–87.
A discussion of current BACI designs, and an enhanced BACI design to detect changes in variability as well as in the mean response.

______. 1994. Things environmental scientists (and statisticians) need to know to receive (and give) better statistical advice. *In* Statistics in ecology and environmental monitoring. D.J. Fletcher and B.F. Manly (editors). Univ. Otago Press, Dunedin, N.Z.

4 RETROSPECTIVE STUDIES

G. JOHN SMITH

Abstract

Statistics are extremely important to resource management. The rigour of gathering and analyzing data for a proper statistical analysis often conflicts with the need to obtain the required information within a short time frame and within a limited budget. Retrospective studies are one alternative to a fully controlled, or prospective, study. These studies offer a compromise, which uses existing data or circumstances. This approach greatly shortens the time between the inception of the study and the presentation of the results, as well as reduces the cost. A considerable degree of methodological correctness can be maintained by careful design, analytical techniques, and presentation of results.

As with any compromise, retrospective studies must be used carefully. In retrospective analyses, often the results are preliminary, and sometimes do not allow for quantitative model building, hypothesis testing, or point estimation. However, by carefully presenting results and designing the study, and being aware of the pitfalls inherent in individual analyses, a great deal of useful information can be obtained. Even if the results are interim, such efforts can be beneficial to the gathering of future information, and to decision-making processes. Managers need all the tools available to properly manage forest resources and adapt to changing conditions and priorities.

In this chapter, a definition and many examples are presented to demonstrate the differences between prospective and retrospective studies. Each example is reviewed with an emphasis on contrasting retrospective and prospective studies, and pointing out the strengths and weaknesses of the retrospective approach. Finally, some suggestions are given regarding the design of retrospective studies and the analysis of retrospective data.

4.1 Introduction

In any study involving data, two values help determine the methodology to apply. The first is *expedience*—to complete the work as quickly and efficiently as possible to meet deadlines and minimize cost. The second is *rigour*—to scrupulously apply statistical methods and experimental controls to ensure that all comparisons and estimates are statistically valid and free of bias and confounding factors.

The values of expedience and rigour usually conflict. However, both are important. In dynamic natural systems such as forests, many effects of events (such as particular logging practices or forest fires) take many years to manifest themselves. Studies have limited time and resources. In the real world, statistics involves art as well as science. To provide valid and accurate results, technical considerations such as experimental or survey design, sample size and allocation, and the desired accuracy or precision of the results are necessary. However, with limited resources, these requirements must be balanced with constraints such as the ability to execute the field procedures, weather, training and management of participants and field personnel, financial resources, and time available to gather and analyze the data.

This chapter discusses some compromises that attempt to satisfy both values. As with any compromise, sometimes it provides excellent results, at other times it is the best of a difficult situation, and occasionally it is unworkable.

Nemec (this volume, Chap. 2) discusses *designed experiments*, those where the experimenter assigns treatments and can manipulate the experimental factors at will. Schwarz (this volume, Chap. 3) discusses the study of *uncontrolled events*, those where the experimenter has a very limited ability to manipulate the experimental factors, and methods for improving the information that can be gained from them. However, what happens when the results of the studies will not be available for a long time, or where it is unrealistic or unacceptable to implement past practices that have been condemned? An example might be a study of the effect of large-scale clearcutting. How can we use data that already exist and what are the advantages and pitfalls?

4.2 Definitions

All studies that involve the gathering and synthesis of data can be placed in one of two categories, depending on the nature of the design, the data, and the analysis. The first category is called a *prospective study*. This term, which has been used extensively in biostatistical literature relating to medical science

(Bailar and Mosteller 1986; Rosner 1986), indicates that the data are collected and analyzed without reference to past data or circumstances.

A prospective study may be either *designed* as described in Chap. 2 or *uncontrolled* (see Chap. 3). The second category is called a *retrospective study*. In this type of study, data that have already been collected for other purposes, or that are of useful historical circumstances, are used directly.

To clarify this point, let us refer to the following definitions from Rosner (1986, p. 326). Although these definitions relate specifically to medical science, the underlying ideas are directly applicable to forestry.

> Definition 10.4: A *prospective study* is a study in which we identify a group of disease-free individuals at one point in time and follow them over a period of time until some of them develop the disease. We then try to relate the development of disease over time to other variables measured at baseline.

> Definition 10.5: A *retrospective study* is a study in which we initially identify two groups of individuals: (1) a group that has the disease under study (the cases) and (2) a group that does not have the disease under study (the controls). We then try to relate their prior health habits to their current disease status.

Although not from forestry, these definitions emphasize conceptual differences between the two approaches for designing a study. In the prospective case, all of the data collection and design of the study are based on events that happen after the inception of the study. In the previous definitions, we start out with similar subjects and then observe what happens. Subjects will be divided into categories based on factors that are suspected to be related to the development of the disease. The researcher will then try to relate the development of the disease to various characteristics and behaviours of the individuals.

In the retrospective case, data that are already available, or are of circumstances or events that have already happened before the study is initiated, are used extensively. In the previous definitions, we start out with two different groups of subjects—one with the disease, the other without—and then work backwards to determine relationships between the acquisition of the disease and the characteristics and behaviours of individuals. The References section contains many references to the statistical treatment of data in particular instances. Most of the statistical literature relates to the medical field (e.g., Bailar and Mosteller 1986; Greenberg 1988; Hoogstraten and Koele 1988; Koele and Hoogstraten 1988; Sikkel 1990; Weinberg et al. 1993).

The distinction between prospective and retrospective is significant, as it represents a great difference in the measure of control over the study's design and execution. Note that this method of categorizing studies does not refer to particular statistical or survey methodologies.

If all things were equal, the obvious choice would be a prospective approach because it provides more control over the design and implementation of the study. A prospective study offers the option to determine exactly what data to obtain, to determine the survey techniques or the design of the experiment, and, in some cases, to exercise control over the types of treatments and their assignment to sampling units.

However, all things are not equal. The use of prior knowledge or taking advantage of existing results is critical in forest management for many reasons. First, enormous savings of time, labour, and financial resources can be realized if data or results from prior studies or surveys are used. Second, in some instances it may be impossible to re-create the exact circumstance that occurred at some previous time or in another location. Third, social and political pressures may prevent the execution of treatments that would likely have significant negative impacts on an ecosystem or community. Retrospective analysis is especially useful in planning and designing future studies. For example, prior studies can help in estimating sample sizes, determining statistical power (Anderson, this volume, Chap. 6) or estimating prior probabilities in a Bayesian analysis (Bergerud and Reed, this volume, Chap. 7). Also, prior data can be used to develop or refine hypotheses or provide information on the required time span of a prospective study.

Another important consideration in the design of a study is the fact that resource managers must make decisions regularly. Because making decisions based on some rather than no knowledge is better, retrospective analysis can be effectively used. A carefully designed retrospective study can be accomplished within a much shorter time frame than a prospective one. The researcher may have less freedom to control the design, but this shortcoming is offset by the requirement for fewer resources and the availability of other sources of knowledge and data. Thus, answers can be obtained quickly from a retrospective analysis, whereas a prospective study might not be completed before a decision must be made.

4.3 Examples

To further illustrate the concepts mentioned in the previous sections, let us consider the following examples.

4.3.1 Example A: effect of landslides on water quality

To assess the impact of major natural disturbances such as landslides on water quality, it would be socially unacceptable to create such events. An alternative is to consider areas where these events have already occurred. Furthermore, significant time, effort, and money are saved by using data from previous natural disasters to compare or observe trends rather than waiting for future events to occur. This analysis would be termed *retrospective* since the data collected will be gathered from events that have already transpired.

This application fits the definition for a retrospective study given in Section 4.2. Let us compare this example with the definition point by point. In the example, we identify two types of areas (corresponding to *groups of individuals* in the medical definition): in one area landslides have already occurred (*cases*), in the other they have not (*controls*). We then relate the water quality (*prior habits*) to the current status of the area. The experimental design will not be as rigorous as for a prospective study because the sampling will be *opportunistic* rather than being based on a statistically rigorous design. To compare the water quality in slide and non-slide areas, it would be ideal to compare areas that were identical in all respects except for the fact that a slide occurred in one and not the other. If this comparison were achievable, then any difference in water quality would be due solely to the slide. However, this comparison will generally not be possible, and hence the analysis and interpretation of the data will need to take this into account.

4.3.2 Example B: forest bird populations

Retrospective analysis is extremely useful in the study of long-term trends. Consider, for example, a study of the long-term effects of clearcutting on forest bird populations. A prospective approach would consist of clearcutting several areas according to a specific design in which natural factors, which affect the regeneration and stand development, can be controlled. The next step would involve collecting data on these areas for the next, say 80 years. Since managers would likely have to make a decision before the results from this study are available, the disadvantage is obvious. Retrospective analysis can help the manager by providing timely information.

An alternative to the prospective approach is to consider areas that have been clearcut in the past. Bird populations can be assessed at various stages of regeneration after clearcutting. This method is a *retrospective* study because advantage is being taken of circumstances that serve as proxies (alternatives) for treatments (in this case, areas that have been clearcut previously). This approach greatly decreases the duration of the study. As in the previous example, a compromise must be made in the experimental design because the sampling will be opportunistic and natural factors such as those previously mentioned would not be controllable to the same degree as in a prospective study.

4.3.3 Example C: economic effect of forest fires

Often starting a study from scratch is difficult. For example, past studies have determined the extent to which long-run timber supply and the flow of economic benefits from a forest can be reduced by fire. To assess the costs and benefits of a fire protection program, the probability of destructive fires must be estimated (Reed 1995) in one of two ways. First, retrospective analysis of historical fire data could be undertaken. The second option, a prospective analysis involving the selection of several areas based on a suitable sampling plan, would involve, according to plan, setting fires in some of the areas, and leaving others undisturbed. The resulting economic effects in the burned areas could be compared to those in areas that were not set ablaze. While this second option is an interesting design, and from a purely statistical point of view has many advantages, it is socially unacceptable. Furthermore, this study would clearly be very long and highly impractical.

4.3.4 Example D: effect of herbicides

The provision of information to direct future studies is an important contribution of retrospective analysis. In 1991, a study investigated the impacts of herbicides on grizzly bear forage production in the

Coastal Western Hemlock zone (Hamilton et al. 1991). Retrospective analysis of past herbicides treatments provided information on the relationship between stand structure, density, and forage availability in glyphosate-treated stands. This information was used to guide the range of treatments to use for further investigation and testing.

4.3.5 Example E: preservation of biodiversity

The idea that green tree retention mimics the stand structure remaining after natural disturbances such as fires is an important assumption that currently guides much of the management for biodiversity (B.C. Ministry of Forests and B.C. Environment 1995). An example of this principle is found in Traut (1995). He studied the preservation of biodiversity in green tree retention—a logging practice where some large trees are left uncut in each cutting unit—by examining areas that had been ravaged by fire. The effects of fire were assumed to be analogous to the effect of green tree retention, because several trees survive the fire as they would survive the logging. Although fire and logging using green tree retention are different, if they have similar effects on biodiversity, effects can be studied that would not be possible otherwise, except by waiting several decades.

Acker et al. (1995) did a similar study in the Willamette National Forest, as did Zenner (1995) on Douglas-fir and western hemlock in the Western Central Oregon Cascades. Peck and McClune (1995) did a similar study in western Oregon, but targeted canopy lichen communities.

4.3.6 Example F1: site index versus tree density

Goudie (1996) presented a paper discussing the relationship between the site index and the tree density in lodgepole pine stands. He postulates that increasing density represses growth, and that to measure the site index on dense stands underestimates the index. It is often assumed that these two factors—density and growth—are independent, thus challenging the usefulness of thinning as a measure to promote growth. The investigation of this phenomenon with a prospective study would require a design in which stands of various densities would be subjected to a variety of thinning regimes, and the effect on the site indices observed.

Goudie extensively used stands that had been thinned in the past, sometimes by natural causes such as fire. He takes great effort to ensure that stands representing various tree densities are similar in all respects except for site density so that a difference in site index can be directly attributable to the density. He correctly points out the potential biases in the selection of plots, and although the design has some shortcomings, he finds compelling evidence to suggest a density-dependent repression on site index. The dependence was most noticeable in very dense stands and hardly noticeable in stands of lower density. Thus growth and yield models that do not take repression into account would not be applicable to very dense stands

Here is an example where existing data have been used considerably. The results were presented along with a discussion of the potential flaws and biases, and a great deal of information and knowledge was gained. Researchers would have had to wait years for the results from a prospective study.

4.3.7 Example F2: site index versus tree density

Thrower (1992) also conducted a study at Larson's Bench, east of Chilliwack, B.C., on the relationship between density and height-and-diameter growth in coastal Douglas-fir stands. This study, too, tried to mimic an experimental design by comparing a natural (unspaced) and a previously logged (spaced) area, each with several similar ecological units. This study used an existing thinning, which was not designed for a research study. Some matching of units was possible. However, since the two areas did not have the same conditions at the time of thinning, it was acknowledged that the comparison of the growth rates in the two areas may be a combination of growth rate and initial conditions.

4.4 Contrasting Data Collected from Prospective and Retrospective Studies

Data collected from a prospective study can be used and analyzed directly, and the interpretation can be based on sound statistical design and analysis. By contrast, data from a retrospective study have fewer statistical controls, and often some components cannot be combined with other components of the data unless additional assumptions are made about their comparability.

Consider Example B (forest bird populations) in the previous section. For a prospective study, the same treatments would be used throughout 80 years of the study. The analysis of trends for various species is relatively straightforward. In the retrospective study there may be a variety of stands where clearcutting was done at various times in the past. The choice of stands will be limited and comparisons may have to be made among stands that differ not only in the time since being clearcut, but also in other variables such as weather, initial stand structure, slope, aspect, and elevation (which cannot be controlled).

In a retrospective analysis, stands must be matched as closely as possible (i.e., choose stands for comparison that are as similar as possible in all ways except for the time since the clearcut). If we can achieve this, then the difference in the populations will be due to clearcutting and not to other factors. Lack of matching requires making additional assumptions about the comparability of the stands included in the analysis. For example, you might assume that for different areas, the weather conditions since the clearcutting will have affected all the clearcuts in the same way. Before making this assumption, a study of each area would be advisable to assess the assumption's plausibility. Thus the lower cost and shorter time of the study are somewhat offset by a design with fewer controls. Because of the greater use of untested assumptions, the interpretation of the results will demand greater caution than in a prospective study. For example, in many instances, results cannot be generalized, nor hypotheses tested. However, hypotheses can usually be generated for future studies.

Retrospective analysis offers other benefits over long-term experiments. In Example B (forest bird populations), a prospective experiment has the risk that some of the treatment areas may eventually be used for purposes other than forests (e.g., become agricultural land or urban areas) during the course of the experiment. Hence the treatment areas are lost to the final analysis. This risk is present even in studies that last for considerably shorter periods of time.

4.5 Comparing the Development of a Retrospective and Prospective Study

A comparison of the development of prospective and retrospective studies, especially where they differ, will help us to appreciate how to use retrospective analysis effectively. Figure 4.1 outlines the basic steps involved in the development of the two types of studies.

The left side of Figure 4.1 follows the steps in a retrospective study, while the right side displays the steps in a prospective study. In a retrospective study, several steps replace the experimental design stage in a prospective study. Comparable data sources must be found that replace rigorous statistical and field procedures. Any difficulty in finding truly comparable data sources inhibits a rigorous statistical design. The lack of control in the retrospective study means that additional assumptions will be required to perform the analysis, implying that more care is needed in the interpretation of the results. For example, if in Example B (forest bird populations) we were unable to match clearcut and natural stands, we may find it necessary to compare a clearcut stand with a natural one that has a different aspect and tree species mix. We would assume that these factors in the two stands do not make a difference to the species present.

In Example F2 (site index versus tree density) the comparability of thinned and unthinned stands was somewhat compromised because of differences in the conditions of the stands at the time of clearcutting. Consequently, the results may be less widely applicable. The advantages, however, are the potential to greatly shorten the study's time frame, and to reduce the effort and resources required.

The previous comments do not imply that retrospective analyses are inadequate, but do indicate the importance of additional diligence during their design, interpretation, and analysis. Furthermore, the study of alternative data sources greatly enhances the ability to design an efficient study using knowledge already gained about the subject under investigation. This study may lead to rejecting some scenarios or considering others that might not otherwise be obvious. Also, retrospective data provide advance warnings of difficulties one might expect, which could lead to the failure of an experimental design. A retrospective study can often be used as a pilot study to obtain qualitative information.

4.6. Studies with Significant Retrospective and Prospective Components

Sometimes it is not clear whether to classify a study as prospective or retrospective. The study may appear to be a prospective study, but after scrutiny may be found to be more like a retrospective one. Thinking about which category the study belongs to will help us understand how the assumptions might affect the interpretation of results. The following examples illustrate this point.

4.6.1 Example G: change in the measuring instrument

Consider the case where a questionnaire survey of

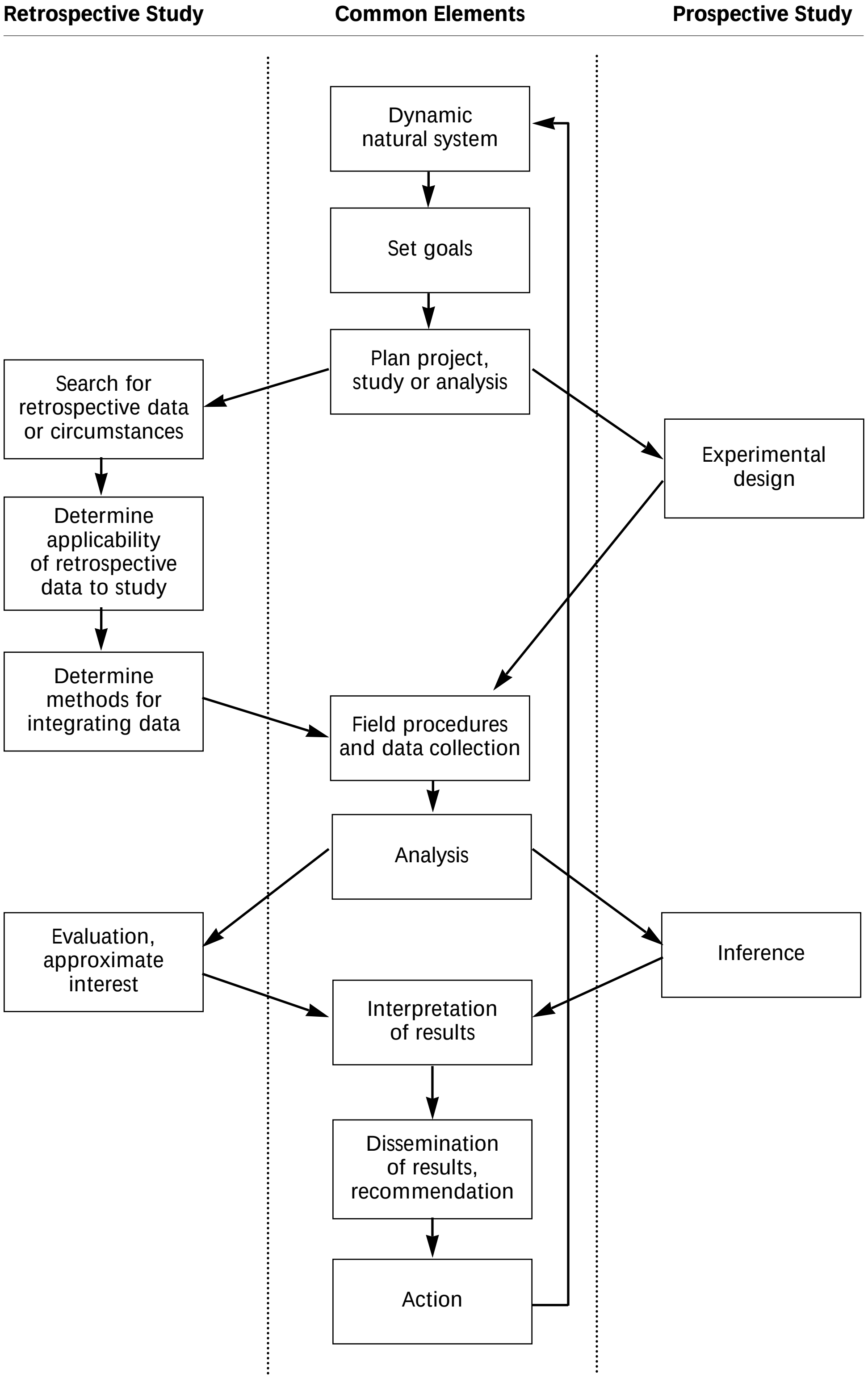

Retrospective Study
Common Elements
Prospective Study
Dynamic natural system
Set goals
Plan project, study or analysis
Search for retrospective data or circumstances
Experimental design
Determine applicability of retrospective data to study
Determine methods for integrating data
Field procedures and data collection
Analysis
Evaluation, approximate interest
Inference
Interpretation of results
Dissemination of results, recommendation
Action

FIGURE 4.1 *Comparing the development of a retrospective and prospective study.*

hunters has been done for many years, yielding information about hunter activity, number of animals or birds killed, etc. With advancing technology and more knowledge of the resource, the survey is redesigned, new computer technology is used, the questions are clearer, and perhaps one or two are added or deleted. The same series of statistics is generated before and after the redesign.

Now suppose that later a trend analysis of the number of days hunted and the number of birds of various species killed is required over a time period that spans both the old and the new methodology. Because the same series of statistics has been generated throughout the period covered by the study, the researcher might infer that it would be valid simply to use the data without further consideration. However, changes in the way a question is asked or in the way the questions are edited may influence a person's response and therefore an apparent trend may be due to the question rather than the activity. Any analysis would require assumptions about the comparability of the data. The interpretation of the results would need to acknowledge these assumptions and examine their potential implication. Cooch et al. (1978) describe some of the effects of survey changes on results in the Canadian Wildlife Service's National Waterfowl Hunter Surveys.

4.6.2 Example H: changing statistical methodology

Consider the following simplified example. Suppose that for a number of years the ratio of immature to adult Ancient Murrelets (*Synthliboramphus antiquus*), called the *I/A ratio*, has been collected for a specific population consisting of five colonies. This ratio is of interest because the higher it is, the greater the number of young per nest. This is one measure of the health of the population. These birds are colonial (i.e., nest close together in small areas or colonies). Each year a sample of approximately 20 nests is observed from each of five colonies under study, and the number of adults and young are counted. The overall I/A ratio was computed by simply averaging the ratios from the five colonies (Table 4.1).

In 1996, the size (i.e., the total number of nests) of each colony, in addition to the number of immatures and adults in the observed nests, is recorded. The results are given in Table 4.2.

As in previous years the comparable I/A ratio can be computed as the simple average of the I/A ratios for the five colonies. Its value is 1.07.

TABLE 4.1 *I/A ratio by colony 1990–1995*

	I/A ratio by colony					Average I/A ratio
Year	A	B	C	D	E	
1990	0.83	1.10	0.90	1.12	1.35	1.06
1991	0.90	1.02	0.84	1.25	1.41	1.08
1992	0.95	0.97	0.88	1.12	1.22	1.03
1993	0.92	1.03	0.81	1.17	1.27	1.04
1994	0.85	0.93	0.78	1.08	1.19	0.97
1995	0.96	1.05	0.91	1.20	1.36	1.10

TABLE 4.2 *I/A ratios for 1996*

Colony	A	B	C	D	E	Total
Nests observed	20	20	18	17	20	95
Immatures	29	39	34	39	63	148
Adults	40	40	36	34	40	190
I/A ratio	0.73	0.98	0.94	1.15	1.58	
Colony size	58	75	105	178	285	701

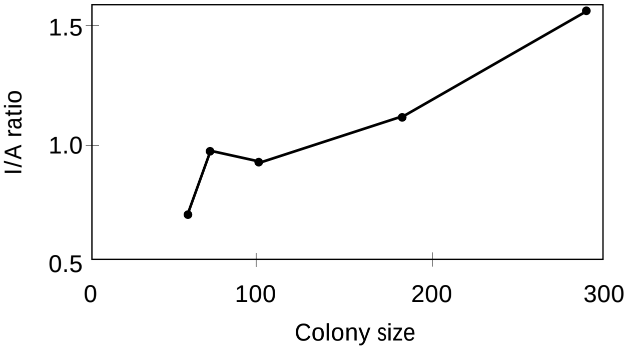


FIGURE 4.2 *I/A ratio vs colony size.*

In 1996, with the information about colony size being added, the I/A ratio can be weighted by the size of the colony. The I/A ratio calculated in this way will be 1.24. In this estimation procedure, the counts are weighted within each colony, which tends to reduce the bias. Furthermore, the I/A ratios are examined as a function of colony size (Figure 4.2).

Subject to confirmation by a statistical test, Figure 4.2 seems to indicate that larger colonies are more productive. The difference between the two estimates of the overall I/A ratio arises because the larger colonies tend to have higher I/A ratios and, in the second estimate, the colonies influence the estimate in proportion to their size.

For management purposes, the "health" of the population in the five colonies is needed. Should the lower figure (I/A=1.07) be presented because it provides the best comparison with information from previous years? Or should the higher figure (I/A=1.24), which will be less biased and a better estimate, be used?

Alternatively, the analysis from previous years could be redone by weighting the results prior to 1996 by the 1996 colony sizes. If colony sizes do not vary a great deal from one year to the next, this method may be a good way to compare as well as update the results from previous years. The downside to this procedure is that if colony sizes fluctuate considerably over time, poorer estimates may result. Furthermore, the agency may present an image of incompetence by not presenting a coherent methodology.

Several solutions may be "correct." The choice will depend on the priorities of the agency and the proposed use of the information obtained. If the data are used as input to a mathematical model, using the best information available would be the highest priority; it may be essential to adjust previous years' information if there is reason to believe this is better. On the other hand, if trends over time are desired, then consistency is important, even at the expense of a systematic bias in each year's results.

The previous two examples represent cases in a continuing study. The researcher has introduced the change in the circumstance through a change in methodology. Often, these cases might not be considered retrospective but simply studies with statistical bias. However, the same factors that exist in Examples A through F, exist here. We are using previously collected data that are not entirely compatible with data we currently collect. The same caveats exist here as in the previous examples. In this sense, Example H is similar to a retrospective study.

The following example demonstrates that a study designed as a prospective one often has elements of a retrospective nature. The existence of such elements should not necessarily result in a study being classified as retrospective. Many studies, of necessity, contain some retrospective and some prospective elements. This happens because we often have limited control over the data we collect. Hence many studies are hybrids and it is an oversimplification to classify every study as either purely prospective or purely retrospective. Consider the following example from wildlife conservation.

4.6.3 Example I: Pacific Brant

Each year, Pacific Brant (*Branta bernicla*) migrate north in the spring and early summer and south in the autumn. During the northward migration, one of the major stopping areas is the Parksville-Qualicum area on the eastern shore of Vancouver Island. The birds stay there for anywhere from several hours to over 10 days before moving on. Proper conservation and management of the Pacific Brant requires knowledge of how many brant use this area, and for how long.

A solution is to count the number of brant in the area, which seems easy because the vast majority feed on the beaches along the seashore. For counts, where, how often, and what data to collect must be determined.

If counts are done at any regular intervals (e.g., daily) then those birds that stay several days may be counted more than once. Furthermore, brant that stay for a shorter time than this interval may be missed altogether. Certainly both situations cannot be accommodated unless other data are used. Multiple counts can be accounted for through the observation of banded birds since the unique band number provides a method for identifying individual birds. This method can work quite well, but simplifying assumptions are necessary. First, assumptions would be required for any prospective banding study, such as the randomness of the banded birds among the population as a whole, and the likelihood of seeing a banded bird (e.g., the band may be hidden if the bird is swimming, or its identification number may not be readable by the observer).

Banding will be expensive unless brant that have already been banded in previous studies can be used. The incorporation of bands from elsewhere introduces a retrospective component into the study and the requirement for additional assumptions. We must consider where the brant were banded. Were brant from some wintering areas subjected to more intensive banding efforts than those from others? Even if we can answer this question, we need to know how many brant from various wintering areas pass through Parksville-Qualicum. Some components of the population may have a high proportion of banded birds, and others may have none. It may be necessary to make some simplifying assumptions about this source of data as well as an assessment of its validity.

The decisions relating to these issues will result in a compromise between optimum statistical methods, which may be impossible to implement, and allowing less desirable retrospective components. The additional assumptions under which the analysis was done should be stated clearly. More information about the Pacific Brant is given in Campbell et al. (1990a), and estimation procedures in Routledge et al. (1998).

4.7 Guidelines for Designing Retrospective Studies

Several examples of retrospective studies, along with some weaknesses and strengths, have been discussed. As practitioners and decision-makers, we need guidance concerning the factors that distinguish a good retrospective study from a mediocre one. Although retrospective and prospective designs and analyses have basic differences, we should endeavour to apply sound statistical principles to both. The main difference is the rigour with which these principles can be applied to each, and to what extent compromises must be made. The following list provides some principles for good design. Although not exhaustive, this list focuses on several points where retrospective and prospective studies tend to differ.

- Probability sampling
- Awareness and clear statement of assumptions
- Design considerations—controlling variation
- Use of direct measurements rather than proxies for the measurements

Let us look at these four principles and consider how we might adhere to these in both retrospective and prospective analyses.

4.7.1 Probability sampling

Probability sampling introduces an element of randomness into the process of selecting the sampling units. As pointed out by Schwarz (this volume, Chap. 3) randomness is essential to virtually all statistical methods. It serves to remove many inadvertent biases and is central to the strict application of statistical theory to experimental design (Nemec, this volume, Chap. 2) and to the proper assessment of inference (Anderson, this volume, Chap. 6).

Probability sampling is often difficult in a prospective study, and is even more arduous in a retrospective one. Consider Example A (effect of landslides on water quality). If we use past data, we have limited choices for our sampling units. For example, to estimate the water quality (chemical composition and concentration of various impurities in the water) 10 years after a landslide, we are limited to the areas where slides have actually occurred and are of the appropriate age. These may not necessarily be representative of the region in which we are interested. For example, if most of the slides were at lower elevations, or in a particular valley, then the data from individual slides may have to be weighted to compensate for their geographical distribution, or a non-random sample be chosen that will have a more representative geographic distribution. This last procedure violates the concept of probability sampling but may represent its best approximation using limited past data.

4.7.2 Awareness and clear statement of assumptions

Often assumptions must be made that are clearly not true, but the consequences of which are hopefully minimal. Consider Example I (Pacific Brant). A retrospective component to this study is the use of bands that were affixed to birds elsewhere.

To simplify the discussion, assume that the brant came from two wintering areas, area X in which a large proportion of brant had been banded, and area

Y in which only a few had been banded. The average length of stay of the brant in Parksville-Qualicum (the first step for estimating the total population passing through the area) can be obtained from repeated band sightings. However, if brant from wintering area X tend to spend more time in the area, our estimate of length of stay will be too high since a disproportionate number of birds seen with bands are from area X. Unless we have detailed banding information and know the proportion of birds from each wintering area, we can do little about this source of error except to assume that the length of stay for the X and Y birds is the same, acknowledge the potential bias, and assess its possible effects on the estimates.

4.7.3 Design considerations—controlling variation

In any study, it is essential to reduce unwanted variation as much as possible. The more we succeed in doing this, the better we can succeed in establishing relationships among variables, testing hypotheses, or obtaining precise point estimates.

In Example F1 (site index versus tree density), Goudie (1996) recognized that stands on which site index were compared must be as similar as possible, leaving tree density as the only variable. His success in determining a valid relationship between site index and tree density depended upon his success in finding nearly identical sites except for the density. Because he did a retrospective study, the job is more difficult—there are fewer stands to choose from. In a prospective study he would have a much larger choice of sampling units—a comprehensive sampling universe from which to sample.

4.7.4 Use of direct measurements rather than proxies for the measurements

Sometimes there are insufficient sampling units with the properties we want to study. In Example E (preservation of biodiversity), this was the case. An assumption was made that a burnt stand with trees left after a fire is equivalent to a logged stand with green tree retention. This assumption allowed a greater choice of stands and thus the potential for better control in the design. However, in doing this, the burnt areas are a proxy for green tree retention. The cost of greater control is the assumption of similarity between burnt and green tree retention stands.

4.8 Roles of Retrospective Analysis in Adaptive Management

Adaptive management is a systematic approach to improving managerial techniques through learning from the outcomes of interventions by those involved with the administration of the resource. This definition implies designing interventions and monitoring programs to provide reliable feedback concerning the outcomes and their causes. Managers need all the tools available to properly manage forest resources, and must be flexible and able to adapt to quickly changing conditions and priorities.

Retrospective analysis can be used to provide input in five important ways:

1. **Assessing long-term management actions without waiting until the effect of the action is realized**. In Example B (forest bird populations), assessing long-term trends in bird populations after clearcutting would require many years. Observing the effect on already clearcut stands can provide useful and expedient information.
2. **Assessing impacts of natural phenomena that cannot be created for the purpose of a study**. Consider Example A (effects of landslides). Landslides are rarely produced on purpose. In cases such as in road construction in mountainous terrain where some slides may be created through blasting, landslides would likely not provide useful information for a study of water quality. A more representative sample could be obtained from natural slides.
3. **Studying historical patterns of events such as disturbance by fire, fluctuations in weather, or outbreaks of parasites such as the gypsy moth**. The study in Example C (economic effect of forest fires) looks at previous patterns of fire and uses these data to assess the economic impacts of the fires. This study also would fit in category 2, above, since fires would not be set simply for estimating impacts, economic or otherwise. Controlled burns for preventing larger fires in the future would not suffice for this type of analysis because they would not furnish a suitable sample for assessing the impacts of large devastating fires.
4. **Collecting background information to aid in the design of a related study**. Example D (effect of herbicides) demonstrates a retrospective study that

was performed to take advantage of existing data to aid in the planning of a more suitable study that will have more focused objectives.

5. **Providing interim information for making decisions when results from long-term studies will not be available until after the intended deadline has passed**. The studies in Example E (preservation of biodiversity) yield considerable information on the relationship between biodiversity and certain logging practices. This information would take many years to collect and would be difficult to implement as a prospective study. Example F1 (site index versus tree density) falls into this category too. A prospective study would take years to complete as the database would build very slowly.

In discussing the role of retrospective studies in adaptive management, their relationship to prospective studies must be considered. In the following discussion, the two methods will be compared and contrasted.

Each of the five ways of using retrospective analysis listed previously can be broken into two components:

1. **Feedback for management**. The opportunity to obtain information for improving current management practices often exists. Regular feedback should not be restricted to retrospective studies. As prospective studies progress, they offer greater potential for feedback, but results typically take a longer time. Wherever possible, prospective studies should report interim results. If this information is prepared with the proper interpretation, the manager will be aware of current ideas, and realize that they are subject to change as more information becomes available. We should not be reluctant to assess the most current knowledge and use it to modify management practices.

2. **Feedback for understanding**. This is just as important as *feedback for management.* In addition to knowing that a particular course of action works, we need to understand why it works. What are the underlying natural processes and relationships that cause the results we observe? It is only when we answer these types of questions that we can generalize our ideas and progress. This type of feedback will be assimilated into the knowledge base for the resource, and while it may not be useful for current decisions, it will become useful in the future when combined with other information.

Feedback for understanding has two components. The first is direct input to the knowledge base. Prospective studies are the best way to achieve this as they are carefully controlled, and statistical accuracy is more important than short execution time. The second component is the provision of direction for future studies. This can often be accomplished by a retrospective study since only general directions are required, and time need not be spent obtaining rigorous results.

4.9 Conclusions

Retrospective analysis is an essential tool for research and management. It allows the researcher to augment and design studies by relying on previous data or circumstances. Its advantages include great savings of time and resources since much work has already been done. In most instances, forest managers cannot afford to wait a long time for results from ideal studies, and often cannot afford to do ideal studies at all. The strength of retrospective analysis lies in its ability to provide an alternative to a purely prospective approach by combining historical data with current information for the production of interim results.

Retrospective analysis is an important predictor for the future. Even though quantitative probabilities (P-values) cannot always be attached to hypotheses, a qualitative understanding of processes or estimation of parameters can be obtained. This knowledge can provide valuable background information and will, at the very least, provide information for future research.

Finally, we should not depend entirely on retrospective studies. Because of their inherent weaknesses, their long-term importance lies in providing information concerning future management and research directions, and pointing out where more detailed prospective studies are necessary. At some point we need a quantitative verification of hypotheses with statistically sound results. Where possible, we must ultimately supplement retrospective studies with prospective ones.

Difficult management decisions of today based on sparse information should become routine decisions tomorrow based on solid information. This transition is ongoing, because with each puzzle we solve, a new one seems to be waiting. Retrospective analysis can be applied to timely but tentative results, followed by primarily prospective studies, and then to thorough investigation of a phenomenon. Our activities must be divided between "fighting fires" and planning for the future in areas not yet under threat.

References

Acker, S.A., P.S. Muir, G.A. Bradshaw, E. Cazares, R.P. Griffiths, G.W. Lienkaemper, B. Marks, B. McCune, A.R. Moldenke, R. Molina, J-L.E. Peck, J. Smith, B.H. Traut, and E.K. Zenner. 1995. Retrospective studies of the effects of green tree retention of conifer production and biodiversity on the Willamette National Forest, Dep. For. Sci., Oreg. State Univ. Suppl. Agreem. PNW 92–0289.

Anderson, J.L. [n.d.]. Errors in inference. This volume.

Bailar, J.C. and F. Mosteller. 1986. Medical uses of statistics, NEJM Books, Waltham, Mass.

Bergerud, W.A. and N.J. Reed. [n.d.]. Bayesian statistical methods. This volume.

British Columbia Ministry of Forests and Environment 1995. Biodiversity guidebook. Victoria, B.C. Forest Practices Code guidebook.

Bunnell, F.L. 1995. Forest-dwelling vertebrate faunas and natural fire regimes in British Columbia: patterns and implications for conservation. Cons. Biol. 9:636–44.

Campbell, R.W., N.K. Dawe, I. McTaggart-Cowan, J.M. Cooper, G.W. Kaiser, and M.C.E. McNall. 1990a. The birds of British Columbia, Vol. I. Royal B.C. Museum, Victoria, B.C.

______. 1990b. The birds of British Columbia, Vol. II. Royal B.C. Museum. Victoria, B.C.

Campbell, R.W., N.K. Dawe, I. McTaggart-Cowan, J.M. Cooper, G.W. Kaiser, M.C.E. McNall, and G.E.J. Smith. 1997. The birds of British Columbia, Vol. III. Univ. B.C. Press, Vancouver, B.C.

Ciampi, A., S. Franceschi, G. Franchini, J. Ghiffault, D. Crivellari, S. Tumolo, and G. Bassignano. 1990. Missing information and confounding in retrospective studies: Statistical methods and an example of data analysis. Statistica Applicata 2:37–51.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. J. Wiley, New York, N.Y.

Cooch, F.G., S. Wendt, G.E.J. Smith, and G. Butler. 1978. The Canada migratory game bird hunting permit and associated surveys. Can. Wildl. Serv. Rep. Ser. 43:8–39.

Goudie, J.W. 1996. The effect of stocking on estimated site index in the Morice, Lakes and Vanderhoof timber supply areas in central British Columbia. Presented at N. Int. Veg. Manage. Ann. Meet. Jan. 24–25, 1996, Smithers, B.C.

Greenberg, R.S. 1988. Retrospective studies (including case-control). Encycl. Statist. Sci. 8:120–4.

Gyug, L.W. and S.P. Bennett. 1995. Bird use of wildlife tree patches 25 years after clearcutting. Rep. prep. for B.C. Min. Environ. Penticton, B.C.

Hahn, G.J. 1980. Retrospective studies versus planned experimentation. Chem. Techn. 10:372–3.

Hamilton, A.N., C.A. Bryden, and C.J. Clement. 1991. Impacts of glyphosate application on grizzly bear forage production in the coastal western hemlock zone. For. Can. and B.C. Min. For. FRDA Rep. No. 165.

Hartman, G.F. and M. Miles. 1995. Evaluation of fish habitat improvement projects in B.C. and recommendations on the development of guidelines for future work. B.C. Min. Environ., Lands and Parks, Fish. Br., Victoria, B.C.

Hoogstraten, J. and P. Koele. 1988. A method for analyzing retrospective pre-test / post-test designs: II. Application. Bull. Psychometric Soc. 26:124–5.

Hunter, M.L. 1990. Wildlife, forests, and forestry—principles of managing forests for biological diversity. Prentice-Hall, Englewood Cliffs, N.J.

Kelsall, J.P., E.S. Telfer, and T.D. Wright. 1977. The effects of fires on the ecology of the boreal forest with particular reference to the Canadian north. Can. Wildl. Serv., Occas. Pap. No. 32.

Koele, P. and J. Hoogstraten. 1988. A method for analyzing retrospective pre-test/post-test designs: I Theory. Bull. Psychometric Soc. 26:51–4.

Lehmkuhl, J.F., L.F. Ruggiero, and P.A. Hall. 1991. Landscape-scale patterns of forest fragmentation and wildlife richness and abundance in the southern Washington Cascade Range. U.S. Dep. Agric., For. Serv., Pac. NW Res. Sta., Gen. Tech. Rep. PNW-GTR-285.

McAllister, M.K. and R.M. Peterman. 1992. Experimental design in the management of fisheries: a review. N. Am. J. Fish. Manage. 12:1–18.

Marcot, B.G. 1989. Putting data, experience and professional judgment to work in making land management decisions. *In* Proc. B.C.–U.S. Dep. Agric. For. Serv. Workshop, Oct. 16–20, 1989, pp. 140–61.

Maymin, Z. and S. Gutmann. 1992. Testing retrospective hypotheses. Can. J. Statistics 20:335–45.

Nemec, A.F.L. [n.d.]. Design of experiments. This volume.

Peck, J.E. and B. McClune. 1995. Remnant trees in relation to canopy lichen communities in western Oregon: a retrospective approach. Dep. Bot. Plant Path., Oreg. State Univ., Corvallis, Oreg. Tech. Rep.

Prentice, R. 1976. Use of the logistic model in retrospective studies. Biometrics 32:599–606.

Reed, W.J. 1995. Estimating the historic probability of stand-replacement fire using the age-class distribution of undisturbed forest. For. Sci. 40:104–19.

Rosner, B.A. 1986. Fundamentals of biostatistics. PWS Publishers, Boston, Mass.

Routledge, R., G.E.J. Smith, L. Sun, N. Dawe, E. Nygren, and J. Sedinger. Biometrics [1998]. Estimating the size of a transient population. Biometrics (in press).

Schwarz, C.J. [n.d.]. Studies of uncontrolled events. This volume.

Sikkel, D. 1990. Retrospective questions and group differences. J. Official Statist. 6:165–77.

Thrower, J.S. 1992. Height and diameter growth of Douglas-fir site trees and Larson's Bench. Rep. to Silv. Br., B.C. Min. For.

Traut, B.H. 1995. Effects of variation in ecosystem carryover on biodiversity and community structure of forest floor bryophytes and understory vascular plants: a retrospective approach. M.Sc. thesis. Dep. For. Sc., Oreg. State Univ., Corvallis, Oreg.

Wang, M.C. 1992. The analysis of retrospectively ascertained data in the presence of reporting delays. J. Am. Statist. Assoc. 87:397–406.

Weinberg, C.R., D.D. Baird, and A.S. Rowland. 1993. Pitfalls inherent in retrospective time-to-event studies: The example of time to pregnancy. Statistics in Medicine 12:867–79.

Zenner, E.K. 1995. Effects of residual trees on growth of young to mature Douglas-fir and western hemlock in the western central Oregon Cascades. M.Sc. thesis. Dep. For. Sci., Oreg. State Univ., Corvallis, Oreg.

5 MEASUREMENTS AND ESTIMATES

RICHARD D. ROUTLEDGE

Abstract

Measurements and estimates are never perfect. The potential for error is always present, especially in field studies on large ecosystems such as British Columbia's forests. In the past, inattention to measurement errors has led to serious management failures, most notably in the related field of fisheries management. This chapter describes fundamental measurement concepts, ways to assess and reduce the potential size of measurement errors, and ways to adjust subsequent analyses and associated management actions to account for these errors.

5.1 Introduction

Do any of these statements describe your views?

> *It is a waste of time to worry about measurement errors. I have enough practice in my field to have reduced measurement errors to a negligible size.*
>
> *If I know that my measurements are not perfect, then I should take several, and average them, maybe throwing out the odd one that is far from the others.*
>
> *I have the resources only to make a subjective guess at the abundance of some minor species. Surely this will be adequate. After all, I am only looking for trends. If the measurement errors are large, and are consistently present, can't we ignore them when we are looking for trends?*
>
> *I don't have to worry about measurement errors. I always take repeated observations and use standard statistical techniques to deal with them. If my measurements do contain large errors, then can't I just take repeated measurements, do a routine statistical analysis, and quote a P-value to silence the pesky biometricians?*
>
> *I have an important job to do. I don't have the time or luxury of worrying about statistical niceties like academics and scientists. I need to get on with managing for forest production.*

If you agree with any of these opinions, then you may find this chapter unsettling.

Adaptive management is a methodology for producing information on the consequences of different management practices. These consequences must eventually be measured, and hence measurement is a key part of any adaptive management project. In general forest management, measurements are taken to evaluate outcomes of management actions, assess trends, and evaluate current states of forest ecosystems. This chapter focuses on measurements of natural resources, such as timber volumes, fish population parameters, or species diversity of communities affected by forest management. The types of measurements that are considered range from counting, to direct physical measurements, to educated guesses.

Careful attention to measurement errors is an essential component of a successful forest management project. In field work, we often deal with quantities that are difficult to measure, and the errors in these measurements are often large. It is tempting to ignore these errors, and manage as though the estimates reflect the true state of the resource. This temptation must be resisted as it could lead to ecological, social, and economic disasters. The collapse of our Atlantic cod (*Gadus morhua*) stocks can be attributed in part to unwarranted confidence in official abundance estimates. Here in British Columbia, inadequate attention to measurement and estimation errors put the Adams River sockeye (*Oncorhynchus nerka*) at risk in 1994.

Even when measurement errors are small, as is often the case in the physical sciences such as astronomy, we eventually push our measurement systems to their limit when we address increasingly difficult questions. If we are not vigilant, management expectations can increase beyond the capacity of the measurement system.

Key components of major projects are often measured by subjective guesses or "eyeball" estimates. Although this approach is often the only feasible way to proceed, such estimates are notoriously vulnerable to large errors that violate the basic requirements of statistical analysis. Without realizing it, we may gradually grow to rely on these "measurements" for more than they can deliver. Furthermore, statistical analyses typically require specific assumptions about measurement errors that are often not satisfied by visual estimates. Close attention to these assumptions will ensure that reliable information necessary for making management decisions is extracted from the data.

It is also easy to be fooled into a false sense of the reliability of one's own measurements. Subtle biases can infiltrate the best measurement protocols. Any measurement processes that provide data for key management decisions need to be checked on an ongoing basis through proper quality control procedures.

This chapter will discuss two types of measurement errors—random and systematic—and how they can be identified and addressed. The assumptions about errors required for standard statistical analyses and the consequences of violating these assumptions will be examined. Specific issues associated with special types of measurements will also be discussed. This chapter focuses on measurement errors themselves. Sampling, the procedure that defines how, when, and where measurements are taken, is discussed in Schwarz (this volume, Chap. 3).

5.2. Measurement Errors

5.2.1 Types of errors

Almost all measurements contain errors. The existence of errors does not imply that the worker has made a mistake. Although careful attention to equipment and protocol may reduce their average size, errors can never be eliminated entirely. Errors come in two types: random and systematic.

For example, repeat measurements of the length of a sawlog will typically differ by small amounts. The differences will be caused by a myriad of small factors, such as placement of the ends of the measuring tape, displacement of the middle of the tape by irregularities in the log, or minor errors in reading the scale. These are random (or chance) errors. If the tape were nylon and had been stretched substantially by extended use, then all the measurements would underestimate the true length of the log. This systematic error or bias often goes undetected.

The following example illustrates the distinction between chance and systematic errors. It also illustrates the need for care even in simple measures such as counting.

Aerial surveys have become a popular way of assessing the size of populations of big game animals. It can be tempting to believe that, from an aircraft, every individual can be seen and counted. Goddard (1967) conducted one of the early assessments of this belief. From extensive field work, he ascertained that 69 black rhinoceros (*Diceros bicornis*) occupied Olduvai Gorge in Tanzania. However, in each of 18 overflights, he never spotted more than 50% of this population, even under ideal conditions. Under poorer conditions, he spotted as few as 35% of the animals. Caughley (1974) reported sighting rates for other African populations ranging from 23% for some African mammals to 89% for cattle. Subsequent studies have confirmed these early findings for a variety of species and habitats.

Goddard's survey results were unpredictable. That is, they were subject to chance errors, attributable to many factors, including type of aircraft, altitude, speed, visibility conditions, observer alertness, and happenstance. For example, if a rhinoceros happened to move just as the observer was scanning that location, it would be more likely to be detected.

Goddard's results also consistently underestimated the true population size. That is, they were subject to systematic errors. In each survey, some animals were not seen. Some animals will always go undetected. They may be hidden under trees, tucked behind cliffs, camouflaged by underbrush, or simply missed because the observer was concentrating on another location, or was fatigued.

The presence of chance errors is relatively easy to detect—repeat the measurement process several times under similar conditions. Different results, with no apparent pattern to the variation, show chance errors at work. In contrast, systematic errors cannot be detected by repeat measurements of an unknown value. Had Goddard just taken repeat counts of a population of unknown size, he could not have detected the bias. Repeated measurements of a known quantity are needed to assess the bias.

5.2.2 Gauging the size of the systematic and chance errors

The bias, or systematic error, in a measurement process is the difference between the mean of an indefinitely long list of measurements and the true value of the quantity being measured. It can be estimated by taking the average of repeated measurements of a known quantity.

Once the bias has been determined, then as long as it does not change, it will be predictable and can therefore be eliminated from the results. In contrast, chance errors are never predictable. At best, we can make some statement about their distribution by inspecting a list of measurements of the same quantity. Deviations from the average of this list will estimate the chance errors. The average size of these deviations (calculated formally through a root mean square) is called the standard deviation.

For a list of measurements, $x_1, x_2, x_3, \ldots x_n$, the average or mean is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

and the standard deviation by

$$SD(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

The standard deviation, $SD(x)$, gauges the average size of the chance errors (i.e., the average amount by which an observation deviates from the average.)

The standard deviation tells us absolutely nothing about the bias. In repeated measurements of a known quantity,

$\bar{x}$ – (the known quantity being measured)

gauges the bias.

5.2.3 The average of repeated measurements and its standard error

By taking repeated measurements the size of the chance errors can be assessed. Averaging these repeated measurements also allows the researcher/observer to reduce their impact. By the very nature of chance errors, some measurements will be positive, and others negative. In the averaging process, some of the positive and negative errors will cancel each other. The remaining chance error in the average will tend to be somewhat smaller than the errors in individual measurements. The standard deviation of an average of *n* independent measurements taken under similar conditions will be decreased through averaging by the factor $\sqrt{n}$. Formally,

$$\mathrm{SD}(\bar{X}) = \frac{\mathrm{SD}(X)}{\sqrt{n}},$$

where $\bar{X}$ represents the average of *n* independently generated values of a variable labelled *X*.

The standard deviation of an average is often referred to as a standard error. The *standard error of a sample average* gauges the average size of the mean's fluctuations from sample to sample. The *sample standard deviation* (Section 5.2.2) gauges the average size of the fluctuations of the values within a sample. In many applications, the sample standard deviation is used solely to estimate the standard error in the sample average.

These two quantities—standard deviation and standard error—provide markedly different information. For example, an adaptive management experiment of the effects of logging on invertebrate stream fauna, might investigate the effects on the abundance of mayfly larvae. The stream bottom will likely have been sampled. The standard deviation of the numbers of mayfly larvae per sampling unit describes the inherent variation in mayfly abundance. Although any effects of forest management practices on this inherent variation may be important, the primary concern is usually about changes in the mean or overall abundance. To this end, the sample average is calculated and its standard error is used to gauge the average size of the chance errors in this estimate.

Does it always make sense to average all the repeated measurements? No. The positive and negative chance errors will tend to cancel each other out unless a single, very large error dominates all the others. Such observations should be singled out for special attention as described in the following section.

5.2.4 Discarding aberrant measurements

From time to time, a data set will contain one or more measurements that look aberrant. They stand apart from the others, and appear as if some gross error was made in taking them. (Statisticians call such values *outliers* in that they lie outside the range of the rest of the values.) It can be very tempting to discard these measurements. Not only do they seem unreasonable, but they also draw attention to possible deficiencies in the measurement process.

Resist the temptation to discard outliers. Instead, take a thorough look at how they were generated. Here are the most likely possibilities.

1. They might just result from an obvious error, such as a misplaced decimal point or inadvertently including counts of other animal species in with the rhinoceroses in the previous example. Such errors should be corrected.
2. Outliers might point to a fascinating and important discovery (e.g., that an anomalously large count of a resident bird population is attributable to a previously unnoticed transient population). Such novel insight will usually be worth exploiting.

3. They might be due to an inherent part of the natural variability that should not be ignored.

If aberrant values are routinely thrown out, the resulting data set will give a false impression of the structure of the system being measured. For example, in a wildlife habitat study, denning habitat might be predicted in an area based on the distribution of the diameter at breast height for a stand. For a stand dominated by small trees, the diameter measurements associated with the few mature trees could be considered as outliers. Discarding these outliers could lead to the conclusion that the stand contains only small-diameter trees incapable of providing valuable denning sites. Management errors could then arise from this erroneous impression.

Furthermore, opportunities for identifying and controlling large sources of error can be lost, and clues to new discoveries may go undetected. For example, a few extraordinarily large-diameter trees in a replanted stand may lead to valuable genetic insight; a pocket of unusually small ones may betray the arrival of a new insect pest.

Outliers must be treated carefully for another reason: they can invalidate standard statistical inference procedures (see section 5.3.1, last paragraph).

5.2.5 Accuracy and precision

A measurement process is said to have high *accuracy* if all errors are typically small. This requires that both the bias and chance errors be small. A measurement process is said to have high *precision* if the chance errors are typically small. Measurements that are precise, but severely biased, are still inaccurate. Figure 5.1 provides a graphical illustration.

Figure 5.1a depicts a high accuracy situation, with small bias and small chance errors. This situation is clearly desirable.

The poor precision in Figure 5.1b can often be improved through careful attention to the sources of chance errors. Goddard's aerial counts, for example, were conducted on a low budget. He essentially hitched a ride on an aircraft flying over the area every chance that came along. At greater cost, he could have arranged for all flights to be conducted with the same aircraft, over a standard flight path, at a constant speed, and under similar visibility conditions. Similar improvements can be achieved in an experimental design context through blocking (see Nemec, this volume, Chap. 2), and in sampling through stratification (see Schwarz, this volume, Chap. 3).

The situation depicted in Figure 5.1c is insidious. The high precision creates a false sense of reliability if the practitioner is unaware of the high bias. Standard statistical analysis techniques do not confront this difficulty. They are designed to deal with chance errors only, not bias. For further technical discussion on accuracy and precision, see Cochran (1977, pp. 15–16).

Another measurement problem arises from modern technology. Digital displays are featured in many

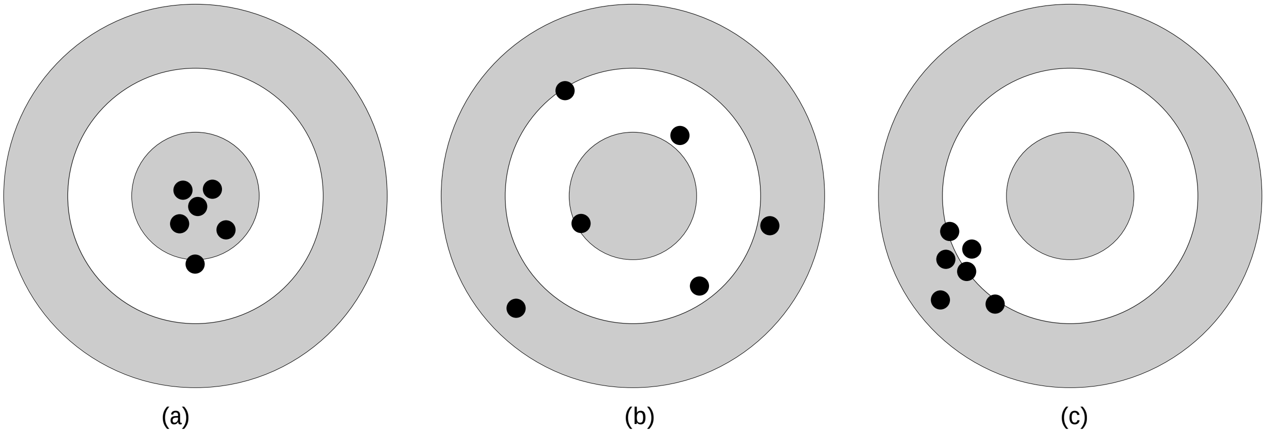


FIGURE 5.1 *Examples of (a) high accuracy (both small bias and small chance errors), (b) low accuracy caused by poor precision (small bias but large chance errors), and (c) low accuracy but high precision (large bias but small chance errors).*

measurement instruments. *It is remarkably easy to forget the inevitable inaccuracy of a measurement, and to accept on faith the accuracy of every digit in an electronic display. When much of the measurement process has been automated, we lose reminders of potential inaccuracy. We must treat a digital readout with as much skepticism as a manual reading of an old-fashioned vernier scale.*

5.3 Standard Statistical Techniques and Measurement Errors

5.3.1 The standard model for measurement errors

In applying standard statistical analyses, measurement errors are assumed to:

1. contain no systematic component;
2. be independent;
3. have a constant standard deviation; and
4. feature a distribution that follows a normal curve.

If any of these model assumptions is violated, then standard statistical analyses may not work properly. Furthermore, as discussed in the next section, a standard regression analysis requires that no appreciable measurement errors be made in the *x*-variable.

Assumption 1 is critical. Consider, for example, a confidence interval designed to capture the true volume of timber in an experimentally managed stand 19 times out of 20. By adding and subtracting roughly two standard errors for the estimate, the formula compensates for chance errors but not for any systematic component. A large systematic error (e.g., generated from a faulty formula for converting diameter and height measurements to volume estimates) would invalidate the confidence interval.

Systematic errors can be insidious. They can grow imperceptibly as equipment ages or observers making subjective judgements change their perspectives. To deal thoroughly with systematic errors, we need to directly assess the bias. If the bias has been determined to a high degree of accuracy, then its value can be subtracted from the measurements. But if the bias estimate is subject to substantial uncertainty, then this factor must be explicitly taken into account in any subsequent analyses. It may be possible to put reasonable bounds on the bias and then manage the resource conservatively.

Bayesian statistics are also becoming popular (Taylor et al. 1997; Bergerud and Reed, this volume, Chap. 7). One might be able to incorporate subjective assessments of the size of the systematic error into a formal Bayesian analysis. Nonetheless, the credibility of this approach will hinge critically on the knowledge base supporting these subjective assessments.

Assumption 2 on independence is also extremely important. This assumption means that in a sequence of independent errors, the value of the next error does not depend on the previous ones. In practice, the error in the next measurement can often be related to the previous one. If, for example, a measurement of nitrate concentration in soil samples were to depend noticeably on the kit purchased to make the measurement, and the next measurement were to be taken with the same kit as the previous one, then the errors would be dependent. If the error in the previous measurement were to be positive, then the next one would likely be positive as well.

In this example, a set of measurements all taken with this same kit would show less variability than a set of measurements each taken with a different kit. Hence, dependent errors can often lead to a false sense of precision. They can often be reduced through randomization (e.g., randomly selecting the measurement kit for each measurement). However, such complete randomization is rarely practical. When other, less complete randomization procedures are followed, the resulting dependent errors can be handled through more advanced analysis techniques such as time-series analysis and analysis of variance for nested or hierarchical designs.

Assumption 3, although often not as critical as the others, can play a crucial role in prediction intervals from regression studies. Prediction intervals in a region of unusually small variability will be wider than necessary, while those in a region of unusually large variability will fail to contain the predicted value with the stated probability.

For example, tree height measurements typically have a nonconstant standard deviation. The standard deviation increases with height. Furthermore, height itself will tend to increase with diameter. In a regression of tree height against diameter, prediction intervals for height would be too wide for small-diameter trees and too narrow for large-diameter trees. (Huang et al. 1992 discuss a common solution to this problem in the context of height-diameter relationships.)

Assumption 4, that the errors have an approximately normal distribution, is not too critical with one important exception—outliers. Mild departures from the normal curve may not seriously alter the behaviour of standard inference procedures, especially if many measurements are being averaged. But watch out for extreme outliers; they can invalidate even the most basic confidence interval or *t*-test. This is another important reason to investigate thoroughly any aberrant-looking measurements (see Section 5.2.4).

5.3.2 Measurement errors in regression analysis

The standard model for regression analysis assumes errors only in the *y*-variable. Errors in the *x*-variable will tend to disperse the scatterplot in the horizontal direction. Consequently, not only does the scatter away from the line increase, but the slope of the regression line also decreases (Figure 5.2). For example, in the management of mixed-species stands such as lodgepole pine and white spruce, a researcher might be interested in establishing a relationship between the site index of white spruce and the site index of lodgepole pine. Site index is an estimated value based on height and age, and is therefore subject to error. The error could be large enough to substantially affect the apparent strength of the regression relationship. Nigh (1995) proposes using the geometric mean regression line in this context. However, this technique, promoted by Ricker (1973) in fisheries analysis, is highly controversial and may itself provide inaccurate slope estimates. The most appropriate technique will depend upon the specific application and on the relative sizes of the variation in the *x*- and *y*-directions. See Fuller (1987) for a thorough discussion of the handling of estimation errors in the *x*-variable.

Outliers are also particularly troublesome in regression analyses. Points that are far from the regression line have considerable influence. Rumours abound of practitioners routinely discarding such points. See Section 5.2.4 on discarding aberrant measurements.

5.4 Types of Measurements and Associated Issues

Following is a discussion of different types of measurements and typical problems associated with each.

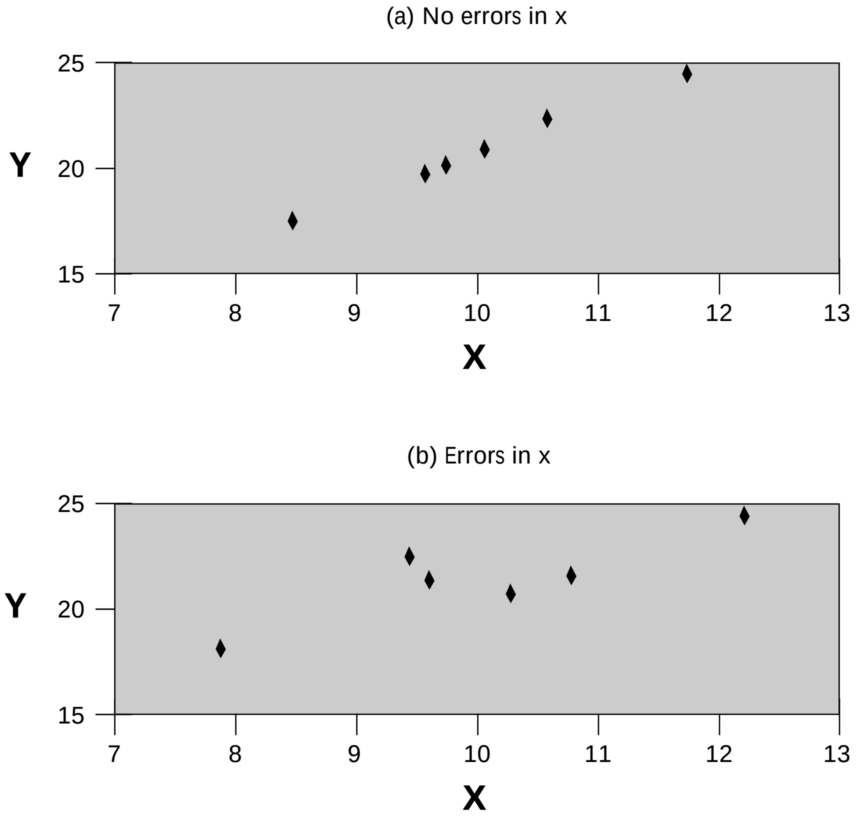


FIGURE 5.2 *Errors in the* x*-values typically not only increase the scatter in the picture, but also spread out the points in the* x*-direction. This result reduces the slope of the regression line. In this example, the slope is reduced from 1.95 to 1.10.*

5.4.1 Counting

Counting is a basic method for assessing the size of a population. As discussed in the example on rhinoceros surveys, it is easy to overestimate the ability to count animals. Be prepared for substantial undercounts. Counts of spawning salmon, for example, can be out by a factor of 5 to 10.

Both over- and undercounting can have serious management implications, and of course these implications extend to other forms of estimates. In a silvicultural management experiment, for example, the numbers of trees in sampled stands may be counted to estimate stand density. If the stand is too dense, then by law it must be thinned; if the stand is understocked, then a remedial measure such as planting must be taken. In this case, over- and underestimation would incur unnecessary cost to the forest manager. The uncertainty in an estimate must always be acknowledged, and quantified where possible through estimates of bias and standard error, along with confidence limits if appropriate.

5.4.2 Direct physical measurements

Measurements of simple physical attributes such as length and mass are usually highly accurate. However, some situations may demand extraordinarily high accuracy. For example, complex models will require extremely accurate measurements if measurement errors are compounded in these models.

5.4.3 Indirect physical measurements

Many measurements of physical quantities are indirect. A liquid thermometer, for example, displays the height of a column of liquid, and measures temperature only insofar as this height is related to temperature. The manufacturer must calibrate the instrument by testing it out at known temperatures. We cannot safely ignore the manufacturer's methodology if we push the instrument close to its limits.

These limits can be exceeded by:

1. demanding more accuracy than can be expected of the instrument (e.g., trying to measure altitude to the nearest metre when repeat readings at the same elevation show chance variation over a range of 20 m);
2. taking measurements outside of the range of values for which the instrument was designed (e.g., using an altimeter calibrated for use up to 3000 m in a flight over Mount Waddington, whose summit is at 4019 m); or
3. using the instrument under unusual conditions not tested in the calibration process (e.g., using an altimeter at extremely cold temperatures, at which it was not tested).

Calibration is not an easy task. The scale on a thermometer has evenly spaced markings, which reflect the approximate straight-line relationship between temperature and volume (and, hence, height of the column of liquid). Many other relationships are curved. Hence, the calibration process involves fitting a nonlinear relationship between actual values and measured responses. Other variables may also influence the relationship. Furthermore, the "actual values" will in fact be measurements themselves. Fuller (1987, pp. 177–179) provides a brief introduction to some of the statistical issues associated with calibration. Also, see the warning in Section 5.3.2 about measurement errors in an *x*-variable used in a regression analysis.

5.4.4 Subjective estimates and trend indicators

Informal management schemes often rely on subjective estimates. Many of British Columbia's smaller salmon populations, for example, have traditionally been subjected to annual, visual surveys from which subjective estimates were made.

Many people regard annual surveys, such as those for British Columbia salmon populations, as useful primarily for monitoring trends. In many instances, reliable indicators of trend are needed instead of unbiased measures of population size. A downward trend in an abundance indicator signals the need for increased protection, perhaps regardless of the actual size of the population.

However, anyone relying on a trend indicator for management decisions must be very careful. Coho salmon (*Oncorhynchus kisutch*) estimation on British Columbia's Central Coast illustrates two of the problems. Coho are notoriously hard to spot, and abundance estimates of the smaller stocks have been very subjective. An observer's subjective guess one year may be strongly influenced by reported estimates from previous years. Hence, a substantial overestimate in one of the early years might easily have become propagated through subsequent years. Such an error would leave an inflated impression of the historic abundance levels. Attempts to rebuild stocks to perceived historic levels would be misplaced.

Furthermore, incomplete records over time can frustrate attempts to determine trends—detecting a

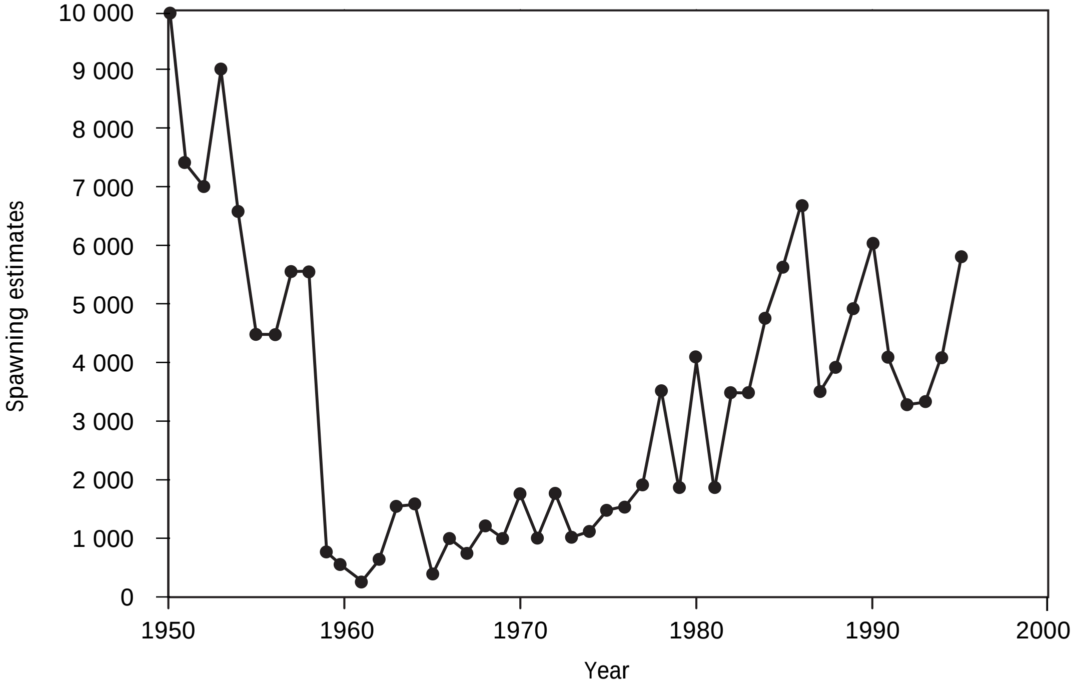


FIGURE 5.3 *Estimated numbers of chinook salmon spawning in the Upper Fraser Watershed near Tête Jaune Cache.*

trend is impossible if no estimates are being collected.

Figure 5.3 illustrates a further problem in using subjective estimates for assessing trends. This graph shows the abundance estimates produced by the Department of Fisheries and Oceans for the chinook salmon (*Oncorhynchus tschawytscha*) that spawn in the Upper Fraser Watershed near Tête Jaune Cache. It looks as if the population has recovered from a near-calamitous decline in the 1960s. One might also argue that the recovery gained strength in the late 1970s, and that abundance reached a short-term plateau around 1980. But the department hired a new local observer in the late 1970s. He proved to be an extraordinarily dedicated employee, who developed a detailed knowledge of the local spawning populations over his early tenure. We shall never know how much of the increase in estimated abundance is attributable to the enhanced knowledge of the resource that this dedicated observer developed.

A rigid adherence to an unambiguous protocol would have eliminated this problem. But we need to foster opportunities for improvement, not stifle them with an inflexible protocol. Furthermore, salmon spawning patterns change over time, sometimes abruptly. The measurement protocol must be free to adapt to these changes. A suggestion for how this transition might be achieved follows.

Often, the conflicting needs for continuity and change can be addressed by developing a new protocol while retaining the old one. Continuity can be maintained by running both protocols in parallel during a test and phase-in period. This continuity also provides an opportunity for fine-tuning the new protocol. Important oversights can be corrected before the new protocol is used for making management decisions.

Problems with observer bias can also be handled through a similar strategy. When observers are changed, the measurement series often changes abruptly. Using multiple observers in a phase-in period can provide a series of parallel observations that can in turn be used to assess the change in observer bias, and provide a training period for the new observer before his/her results are used for making key management decisions. If thorough attention is paid to the measurement protocol, then a regression or time-series analysis can provide more definitive evidence of a real trend.

5.4.5 Turning indices into unbiased estimates

In managing fish or waterfowl populations, we can sometimes reduce reporting biases to obtain reasonably accurate estimates of the numbers killed by humans. However, we often want to know the fraction killed. To do this, we need an unbiased estimate of abundance. By contrast, it is often feasible only to obtain an index of abundance for much of the population. Indices can be turned into unbiased estimates through a sort of calibration process.

In resource management work, this goal is often achieved through double sampling and ratio estimation. The Canadian Wildlife Service, for example, conducts an annual survey of breeding waterfowl. Aerial surveys are used to obtain rough abundance estimates. These figures are then supplemented by more thorough ground surveys. The results of the ground surveys are used to adjust the less accurate, but more extensive, aerial surveys for bias. In a sense, the ground surveys are used to calibrate the aerial surveys.

This technique of double sampling is described in more detail by Cochran (1977, Chapter 12) and Thompson (1992, Chapter 14). It is a valuable tool in a wide variety of contexts. For example, in a management experiment involving the monitoring of grass biomass over time, definitive estimates of biomass could be obtained only through destructive sampling. By contrast, extensive but imprecise information could be obtained from subjective, visual estimates. By sampling a small number of quadrats destructively, the results of a more extensive set of visual estimates can be adjusted for bias.

5.4.6 Quantitative measures of imprecise concepts

Forest management is closely linked to ecology, which in turn is full of vague concepts such as niche width, niche overlap, similarity, importance, competition, and diversity (Krebs 1994). Developing precise, quantitative measures of these concepts is one of the enduring challenges of the subject. The following discussion illustrates both the need for precise, quantitative measures and common problems encountered in their construction and use. These are illustrated in the context of diversity measures.

Biodiversity has recently received increasing attention in resource management. Yet it is not easy to define and measure. The Biodiversity Guidebook in the Forest Practices Code (B.C. Ministry of Forests and B.C. Environment 1995) defines the concept as follows:

> Biological diversity (or biodiversity) is the diversity of plants, animals and other living organisms in all their forms and levels of organization, and includes the diversity of genes, species and ecosystems, as well as the evolutionary and functional processes that link them.

This definition explicitly mentions the organisms and levels of organization to be considered, but leaves the word “diversity” undefined. The guidebook gives directions on how to manage forests to maintain biodiversity. Some of these directions are based on assumptions on how forest ecosystems function, including:

> The more that managed forests resemble the forests that were established from natural disturbances, the greater the probability that all native species and ecological processes will be maintained.

Although this assumption has considerable intuitive and practical appeal, how can we be sure that it is valid and that the strategy is working? We need to have some way of quantifying biodiversity so that we can monitor it directly. This in turn requires a quantitative definition that pins down this vague concept.

No single definition will be universally applicable. A wildlife biologist will be interested in maintaining wildlife diversity by maintaining the structural diversity in a forest. Thus, two concepts of diversity are invoked: the species diversity of the wildlife and the structural diversity of the forest. A fisheries biologist will focus on maintaining the diversity of individual fish stocks (not species), which in turn depends on maintaining a diverse set of healthy fish habitats.

Indices are useful measures of abstract concepts. However, a single measure may not capture the concept fully, and several different types of indices may be needed to measure an imprecise concept. Consider an analogy of blind people trying to describe an elephant. Each person examines a different part of the elephant. No one person will obtain an accurate overall impression of the elephant. Obviously, the more of the elephant that you can include in the operational definition the better, but there will always be limitations.

Now consider species diversity: its simplest definition is the number of species. But counting or estimating the number of species in a community is very difficult. An indefinite number of rare or cryptic species may go undetected in any survey.

Furthermore, diversity depends not only on the number of species, but also on the lack of predominance of any one species. Few people would view a cornfield with a few scattered weeds from seven other species as being as diverse as an alpine meadow containing a more even mixture of eight species. Alfred Russel Wallace's (1895) description of the diversity in the Amazon rainforest captures the essence of the concept:

> If the traveller notices a particular species and wishes to find more like it, he may turn his eyes in vain in any direction. Trees of varied forms, dimensions, and colours are around him, but he rarely sees any one of them repeated. Time after time he goes towards a tree which looks like the one he seeks, but a closer examination proves it to be distinct.

Numerous researchers have attempted to develop a quantitative measure of diversity that would embody such an impression—a measure that both captures the essence of the concept and is easy to implement. To date, most attention has been focused on three measures of diversity (Routledge 1979). All but the first require data on relative abundances of the species. These measures are defined as follows (with $p_1, p_2, p_3, ..., p_s$ representing the proportion of the total abundance contributed by species 1, 2, 3, …,s):

1. the species richness,

$$N_0 = s\,;$$

2. an index related to the Shannon-Wiener index from information theory,

$$N_1 = \exp\left[-\sum_{i=1}^{s} p_i \, 1\mathrm{n}(p_i)\right]; \text{ and}$$

3. an index related to an early proposal by Simpson,

$$N_2 = \left[\sum_{i=1}^{s} p_i^2\right]^{-1}$$

In each instance, if all species are equally abundant, the diversity reduces to the species count, $N_{0.}$ It is customary to view N_1 and N_2 as describing the number of equally common species that would produce an equivalent impression of diversity.

Which index should be used? A choice of index should in general depend on the properties of the index in relation to the goals of a study. The example portrayed in Figure 5.4 and Table 5.1 illustrates some of the differences amongst these three indices. Figure 5.4 displays the abundance patterns; Table 5.1, the values of the diversity indices.

Note how the species richness is unaffected by the relative abundances. Of the other two indices, Simpson's index is less sensitive to the rarer species. Compare, for example, abundance patterns (c) and (d). As the rarer species diminish in abundance, the Shannon-Wiener index declines from 2.19 to 1.34 (i.e., by 0.85). By contrast, Simpson's index declines by only 0.42. If, from (d), the three rarest species were to go extinct, and the abundances of the other five were to remain unchanged, then the Shannon-Wiener index would decline by a further 7% versus only 2% for Simpson's index. Thus, the Shannon-Wiener index is more sensitive to the abundance of the rarer species.

If it is important to focus on the rarer species, then Simpson's index may be inappropriate. But the sensitivity of the Shannon-Wiener index to the rarer species comes at a cost. It is usually difficult to assess the abundance of rare species, and the sensitivity of the Shannon-Wiener index to these quantities makes the index very hard to estimate (Routledge 1980). To date, no generally reliable method has been found to estimate this index.

When choosing the suite of quantities to be measured, we need to remember the specific goals of the project. Simpson's index, for example, may play a useful role in assessing trends in dominance of the more abundant species, but is inadequate in management projects where the preservation of rarer species is a priority. If focusing on the rarer species is important, the Shannon-Wiener index would be more useful if only its bias and standard error could be more readily predicted. A sensible choice would be to estimate Simpson's index along with the abundances of important rare species.

When using an index measure in an analysis, we must be aware of the range and scale of the index so as to make proper interpretations. If, for example, a diversity index were to drop from 2 to 1, should we be concerned? If the diversity were measured by either N_1 or N_2, this could be interpreted as follows: What was once comparable to a community of two equally abundant species has dropped essentially to a monoculture. Such a decrease would be clearly noticeable, and worthy of attention. Had diversity been measured by $\ln(N_1)$, as is sometimes suggested, then a decline from 2 to 1 would have been hard to interpret.

Choosing an index to quantify an imprecise concept is tricky. One must pick an aspect of the elephant that is relatively easy to measure and relevant to the purpose of the study. Overall weight may

TABLE 5.1 *Diversity measures for the abundance patterns in Figure 5.4*

Community	Species richness (N_0)	Shannon-Wiener (N_1)	Simpson's (N_2)
a	8	8	8
b	8	4.81	4.38
c	8	2.19	1.53
d	8	1.34	1.11

be hard to measure; however, a simple measure of height may serve as a useful proxy. Following is a set of guidelines for selecting and using indices.

5.4.7 Guidelines for selecting and using indices

1. Determine the goals of the experiment.
2. Define, as clearly as possible, the abstract concept that is to be assessed.
3. Keeping in mind the goals of the experiment, decide what aspect(s) of the underlying concept are important in terms of the management or scientific objectives.
4. Identify indices that have been developed to quantify this abstract concept.
5. Assess the sensitivity of the indices to the important aspects identified in step 3.
6. Examine the range and scale of the indices. Is it easy to tell whether an observed value represents a desirable level, or whether an observed change in the index represents an important change in the achievement of management objectives?
7. Ensure that the bias and standard error of all selected indices are well understood and predictable. Each index shoul not be overly sensitive to small

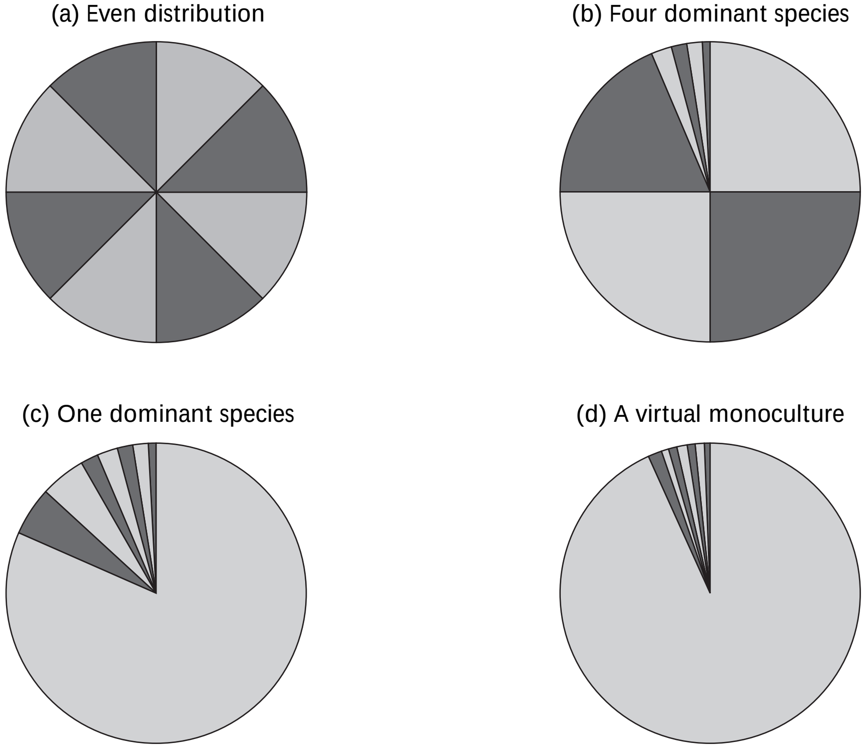


FIGURE 5.4 *Four dominance patterns. In each instance, eight species are present, but the communities are increasingly dominated by fewer species.*

measurement errors. Large, unquantifiable biases are particularly troublesome and can easily arise if the index depends on complex mathematical functions of the measured values, but can also arise in simple indices, such as the species richness.

5.5 Reducing Measurement Errors

Although some measurement errors are inevitable, they can often be reduced substantially. The resulting benefits to the experiment can be considerable. Following are some guidelines for achieving this goal.

5.5.1 Guidelines for reducing measurement errors

1. Counting
   - Ensure that new personnel are trained by experienced workers.
   - Where feasible, mark or discard items previously counted to reduce double-counting.
   - Anticipate undercounting. Try to assess its extent by taking counts of populations of known size.
   - Try to reduce errors by taking counts only in favourable conditions and by implementing a rigorous protocol.
2. Physical measurements
   - Instruments should be calibrated before first use, and periodically thereafter.
   - Personnel should be trained in the use of all measuring devices.
   - Experienced personnel, as part of an overall quality control program, should spotcheck measurements, particularly those taken by new personnel.
   - Incorporate new equipment where appropriate (e.g., lasers and ultrasound, for distance measurements).
3. Remeasurement
   - Watch for the transfer of errors from previous measurements (e.g., a mistaken birth from an item erroneously marked as dead).
   - Reduce errors in relocating the site of previous measurements through more careful marking, use of modern electronic GPS technology, etc.
   - Ensure that bias is not propagated through the use of previous measurements as guides to subsequent ones. (This issue is particularly troublesome in subjective estimates.)
4. Visual estimates
   - Ensure that all visual estimates are conducted according to rigorous protocols by well-trained observers.
   - Pay particular attention to observer bias. When bringing a new observer into the program, ensure that his/her results are backed up by an experienced observer's.
   - If sites or times are to be selected as part of the collection of visual estimates, eliminate selection bias by providing a protocol for site- or time-selection. Do not, for example, let vegetation samplers pick "modal" sites.
5. Data handling
   - Record data directly into electronic form where possible.
   - Back up all data frequently.
   - Use electronic data screening programs to search for aberrant measurements that might be due to a data handling error.
   - Design any manual data-recording forms and electronic data-entry interfaces to minimize data-entry errors. In the forms, include a field for comments, encourage its use, and ensure that the comments are not lost or ignored.

5.6 Summary

A century ago, British Columbia's renewable resources seemed so limitless that we asked very little of our measurements of the resources and their support systems. With new requirements imposed (e.g., by the Forest Practices Code) and with increased harvesting capacity, we are escalating our demands on the measurement systems. In recent years, our systems for estimating fish populations have let us down. The recent controversy over the management of Fraser River sockeye (Fraser et al. 1995), has not been so much about a breakdown in the quality of the measurement procedures, as about the fact that our management expectations have increased beyond the capacity of the measurement system.

Assess continually the adequacy of a measurement system to improve it where possible and to point out when its limitations may be exceeded. The assessment should include:

- the choice of quantities to be measured;
- the procedures and equipment for taking the measurements;
- any associated sampling;

- the processing, storage, and analysis of the data; and
- the demands and expectations of the resource managers or others who use the results.

Good quality control procedures are an essential component of any measurement process. This applies even if the measurements are to be used solely for indicating trends. A well-designed protocol must be followed, and the reliability of the data must be commensurate with management needs.

Just as expectations can rise without notice, so can a gradual deterioration in quality go undetected. Consider, for example, the monitoring of spawning habitat in Devoe Creek far up the North Arm of Quesnel Lake. The creek enters the lake through an old western redcedar (*Thuja plicata*) grove, containing many downed trees and rampant with devil's club (*Oplopanax horridus*). Workers faced with a rising wind on the long stretch of water back to the landing would be tempted to cut corners if they believed that no one valued their work or would ever check on their accuracy. For field measurements, often taken in isolated conditions, quality checks will inevitably be infrequent. *The quality of the data will depend critically on the reliability and commitment of the field staff; sound, quality management practices will foster the required spirit.*

The myths presented at the beginning of the chapter should now have been dispelled.

> *It is a waste of time to worry about measurement errors. I have enough practice in my field to have reduced measurement errors to a negligible size.*

Measurement errors in many field studies are large, and inadequate attention to them has led to major management disasters.

> *If I know that my measurements are not perfect, then I should take several, and average them, maybe throwing out the odd one that is far from the others.*

Taking repeated measurements allows the researcher to assess the average size of the chance errors. Averaging these measurements will usually reduce the impact of the chance errors. However, aberrant measurements should be singled out for special attention, not casually or routinely discarded. They could provide valuable insight, and are an important part of the information collected. In addition, averaging will not reduce any systematic bias.

> *I have the resources only to make a subjective guess at the abundance of some minor species. Surely this will be adequate. After all, I am only looking for trends. If the measurement errors are large, and are consistently present, can't we ignore them when we are looking for trends?*

We need to know enough about the errors to be able to distinguish between a trend in the quantity being measured and in the measurement errors. Furthermore, a false estimate of the historical state of the forests could lead to inappropriate management actions.

Trend indicators are often set up when it seems too difficult or costly to implement the rigorous procedures required to produce unbiased abundance estimates. Trend indicators demand almost as much rigour, and measurement procedures must be rigorous enough to rule out any cause for a trend other than a change in abundance.

> *I don't have to worry about measurement errors. I always take repeated observations and use standard statistical techniques to deal with them. If my measurements do contain large chance errors, then can't I just take repeated measurements, do a routine statistical analysis, and quote a p-value to silence the pesky biometricians?*

The standard statistical analysis procedures require specific assumptions about the measurement errors. Violated assumptions lead to questionable analyses and management decisions.

> *I have an important job to do. I don't have the time or luxury of worrying about statistical niceties like academics and scientists. I need to get on with managing for forest production.*

Adaptive management of British Columbia's forests is an important task. Thorough attention to measurement errors and other statistical niceties will help, not hinder, the ongoing development of improved management strategies.

5.6.1 Guidelines for developing a measurement protocol

Following are recommended guidelines for improving the quality and value of measurements. As with any general guidelines, these will need to be adapted to specific applications, and are intended more as an initial checklist of important items to consider than as an inflexible set of rules.

1. *Determine what parameter needs to be measured. In so doing, pay attention both to the relevance of the quantity being measured and the practical difficulties in obtaining reliable measurements or estimates. Consider both the concept and its measurement.*
2. *Determine the demands that need to be placed on the quality of the resulting measurements or estimates.*
3. *Devise a measurement system that will meet these demands. If this task is impossible, reassess the management objectives and strategies and revisit guidelines 1 and 2.*
4. *Assess the accuracy of the proposed measurement system by taking repeated measurements of known quantities under a variety of conditions.*
5. *Establish an unambiguous protocol for taking measurements, and ensure its proper implementation.*
6. *Implement a system of periodic checks on the continuing performance of the measurement system in light of internal changes and external demands. Watch for subtle increases in the demands placed on the system.*
7. *Look for ways to develop incremental improvements while maintaining the integrity of any long-range data series. When implementing changes, phase them in, running new and old methods in parallel during a transition period.*

References

Bergerud, W.A. and W.J. Reed. [n.d.]. Bayesian statistical methods. This volume.

British Columbia Ministry of Forests and B.C. Environment. 1995. Biodiversity guidebook. Victoria, B.C. Forest Practices Code guidebook.

Caughley, G. 1974. Bias in aerial survey. J. Wildl. Manage. 36:135–40.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. J. Wiley, New York, N.Y.

Fuller, W.A. 1987. Measurement error models. J. Wiley, New York, N.Y.

Goddard, J. 1967. The validity of censusing black rhinoceros populations from the air. East Afr. Wildl. J. 5:18–23.

Huang, S., S.J. Titus, and D.P. Wiens. 1992. Comparison of nonlinear height-diameter functions for major Alberta tree species. Can. J. For. Res. 22:1297–304.

Krebs, C.J. 1994. Ecology: The experimental analysis of distribution and abundance. 4th ed. Harper and Collins, New York, N.Y.

Nemec, A.F.L. [n.d.]. Design of experiments. This volume.

Nigh, G.D. 1995. The geometric mean regression line: a method for developing site index conversion equations for species in mixed stands. For. Sci. 41:84–98.

Ricker, W.E. 1973. Linear regressions in fisheries research. J. Fish. Res. Board Can. 30:409–34.

Routledge, R.D. 1979. Diversity indices: Which ones are admissible? J. Theoret. Biol. 76:503–15.

______. 1980. Bias in estimating the diversity of large, uncensused communities. Ecology 61:276–81.

Schwarz, C.J. [n.d.]. Studies of uncontrolled events. This volume.

Taylor, B., L. Kremsater, and R. Ellis. 1997. Adaptive management of forests in British Columbia. B.C. Min. For., For. Practices Br., Victoria, B.C.

Thompson, S.K. 1992. Sampling. Wiley-Interscience, New York, N.Y.

Wallace, A.R. 1895. A narrative of travels on the Amazon and Rio Negro: with an account on the native tribes, and observations on the climate, geology, and natural history of the Amazon Valley. Greenwood Press, New York, N.Y. Reprinted 1969.

6 ERRORS OF INFERENCE

JUDITH L. ANDERSON

Abstract

Occasional erroneous conclusions (errors of inference) are unavoidable in the analysis of results from management experiments and monitoring programs. However, their probability of occurrence in a given experiment can be controlled. In an experiment comparing two treatments, conclusions can be incorrect in two ways: (1) concluding that a difference between the treatments is real when in fact it is not (a "Type I" error), or (2) concluding that there is no difference between treatments when in fact a difference exists (a "Type II" error). Both types of error can be costly in typical adaptive management experiments, where treatments involve the effects of commercial activities, such as harvesting, on ecosystems. A Type I error may lead to unnecessary limitations on commercial activities, while a Type II error may result in the continuation of activities damaging to the ecosystem. Type I error is limited by the conventional significance level of statistical tests to a frequency of less than five errors per 100 tests performed. The method for estimating and limiting Type II error rate ("statistical power analysis") is less well known but just as important. This chapter discusses conceptual and practical aspects of statistical power analysis (including references and software that aid in performing power analysis) and its role in the design of large-scale experiments in forest management.

6.1 Introduction: Type I and Type II errors

A major goal in adaptive and experimental management is to improve our understanding of managed biological systems by making reliable conclusions (inferences) from experiments and monitoring programs. However, any experimental inference has a chance of being incorrect, and these errors can result in large economic and ecological costs. Therefore, experimenters must understand how errors of inference occur and how to control them. This chapter discusses the following topics:

- the relationship between the two types of errors of inference in statistical tests, with a focus on the category of error most often ignored—Type II error, failure to reject a null hypothesis when it is in fact false;
- factors that influence the probability of a Type II error;
- strategies by which those factors can be manipulated in experimental design to control error rates in ecological studies; and
- software and literature dealing with theoretical and practical aspects of Type II error and its management.

6.2 What Are Type I and Type II Errors, and How Do They Fit into Statistical Inference?

6.2.1 Statistical inference is an important part of the process of evaluating a scientific hypothesis

Statistical inference is a form of reasoning that leads to rational conclusions about states of nature when the available information comes from a sample of the population or system under study (Kirk 1982). Figure 6.1 illustrates the position of statistical inference in the iterative process of evaluating a *scientific hypothesis*. Usually the first step is the statement of the hypothesis as a testable proposition that is tentatively adopted as an explanation for observed facts and as a guide for investigation (Kirk 1982). For example, an ecologist studying the effect of logging practices on Coeur d'Alene salamanders near S4

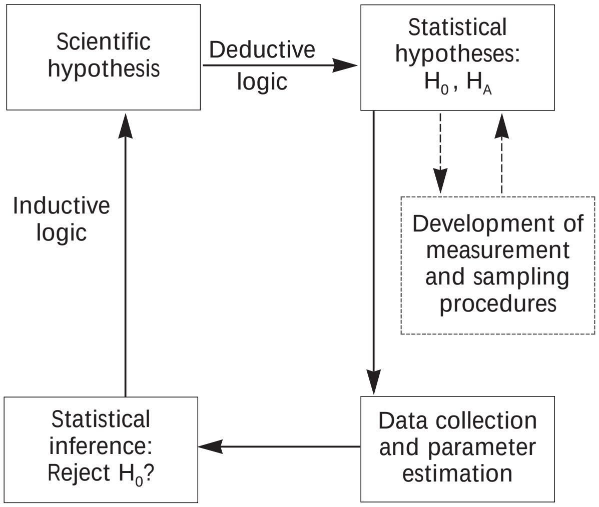

Scientific hypothesis
Deductive logic
Statistical hypotheses: H_0, H_A
Development of measurement and sampling procedures
Inductive logic
Statistical inference: Reject H_0?
Data collection and parameter estimation

FIGURE 6.1 *The relationship between scientific and statistical hypotheses (adapted from Kirk 1982). H_O denotes the null hypothesis, H_A denotes the alternative hypothesis.*

streams in the Kootenay Valley might begin with the scientific hypothesis: "The 10 m wide riparian reserve zone recommended by the Forest Practices Code (B.C. Ministry of Forests and B.C. Environment 1995) is sufficient to maintain salamander populations as a component of vertebrate biodiversity."

Next, deductive logic is used to generate a *statistical hypothesis* that follows from the scientific hypothesis. This statistical hypothesis is a quantitative statement about some variable whose sampling distribution can be described statistically. A statistical hypothesis that proceeds logically from the salamander example might then be as follows: "Population density of salamanders is not reduced in experimental 10 m wide riparian reserves on S4 streams, compared with the density of control populations in comparable habitat in unlogged watersheds."

After identifying the experimental units, the experimenter decides on the statistical parameter to be estimated and develops appropriate measurement and sampling procedures. In the salamander example, the experimental units would be watersheds and the estimated parameter would be mean population density across sampled watersheds in two treatment groups: an experimental group of logged watersheds with riparian reserves and a control group of unlogged watersheds. Further refinement would transform the statistical hypothesis into mutually exclusive null and alternative hypotheses, denoted H_O and H_A, respectively. Examples might be "H_O: mean population density in the experimental riparian reserves (μ_E) ≥ mean control population density (μ_C)," and "H_A: $\mu_E < \mu_C$."

Statistical inference provides a logical structure for drawing conclusions about whether the null hypothesis can be rejected. The basis for this decision is the statistical test, which uses a statistical model to assess the probability that the observed data could have come from a population of experimental units for which the null hypothesis is true. If the probability is sufficiently low, we reject the null hypothesis H_O and, by implication, provisionally accept the alternative hypothesis H_A. This decision inductively influences our degree of belief in the scientific hypothesis under consideration. Typically, in basic science, several iterations of this process are required before consensus is reached about the plausibility of the scientific hypothesis.

In the salamander example, if the mean of population density measurements in watersheds with riparian reserves were significantly lower than that of controls, statistical inference would reject the null hypothesis. This experiment alone would provide one line of evidence about the scientific hypothesis, suggesting that the recommended riparian reserves may not be sufficient to preserve the salamander population as an element of biodiversity. If making a decision about the adequacy of this management recommendation was not urgent, further studies would be desirable to provide a more thorough test of the scientific hypothesis.

6.2.2 Statistical inference can lead to any one of four outcomes, two of which are incorrect

The four possible outcomes of a statistical hypothesis test are shown in Table 6.1. If the null hypothesis is really true, then two outcomes are possible: Not rejecting H_O is a correct inference, while rejecting it constitutes a Type I error. Similarly, if H_O is really false, the correct inference is to reject it, and failing to do so constitutes a Type II error. The probability of committing a Type I error is α, while the probability of a Type II error is β. The power of the test ($1-\beta$) is the probability of correctly rejecting H_O when it is really false.

In the salamander example, a Type I error would

TABLE 6.1 *Four possible outcomes of a statistical test of a null hypothesis. The probability of each outcome is given in parentheses. Management decisions that might proceed from the inference are indicated in parentheses. Adapted from Toft and Shea (1983).*

	Inference	
State of nature	Do not reject H_O (Manage as though H_O were true)	Reject H_O (Manage as though H_A were true)
H_O true	Correct ($1-\alpha$): Correctly infer that no treatment effect exists	Type I error (α): Infer that treatment effect exists when in fact there is none
H_O false	Type II error (β): Infer that no treatment effect exists when in fact there is one	Correct (power = $1-\beta$): Correctly infer that treatment effect exists

be committed if the experimenter concluded that the density of experimental populations had decreased when in fact they were really doing at least as well as the controls. If management decisions followed from this erroneous inference, the forest industry might unnecessarily be required to leave a wider reserve along S4 streams.

On the other hand, a Type II error would be committed if the experimental population had really declined relative to controls, but the experiment failed to detect the decline. In this case, managers would conclude that 10-m riparian reserve zones were sufficient to maintain the species. They might implement this level of riparian protection on a large scale, without realizing that the practice, in fact, failed to meet the objective of maintaining the salamander population in the community.

6.2.3 For a complete analysis of possible errors of inference, the alternative hypothesis must indicate a treatment effect worth detecting

In statistical tests, the data actually observed in an experiment are compared with the data that could be observed if the population under study had a particular assumed value for one of its statistical parameters. To represent that assumed value, a null hypothesis must always include an equality relationship (often called a "point null hypothesis"). It may include an inequality as well if treatment effects are expected to occur in one direction. Thus, H_O: $\mu_1 \geq \mu_2$ and H_O: $\mu_1 = \mu_2$ are valid null hypotheses, but H_O: $\mu_1 < \mu_2$ is not.

In the salamander example, only the equality relationship of the null hypothesis (H_O: $\mu_E \geq \mu_C$) is actually used in the statistical test. The experimenter calculates the observed value of the *t*-statistic $(\overline{x_E} - \overline{x_C})/s$, where $\overline{x_E}$ is the mean of population density estimates in experimental (logged with 10-m riparian reserve zones) watersheds; $\overline{x_C}$ is the mean of population density estimates in control (unlogged) watersheds; and *s* is the standard error of the difference between the sample means. This observed value of *t* is compared with the distribution of estimates of *t* that could arise if the salamander population densities really were equal in both the experimental and control watersheds. H_O is rejected if the calculated *t* value is extreme, such that the probability that it could come from that t distribution is less than the "critical α" (the maximum acceptable rate of Type I error, set by convention to 0.05). In other words, if H_O were really true and the experiment were repeated many times, the null hypothesis would be rejected (incorrectly) in at most 5% of the replicate experiments.

Because traditional practice has focused only on rejection of the null hypothesis, scientists have usually been content to use alternative hypotheses with no equality relationship, such as H_A: $\mu_1 < \mu_2$. However, to analyze the probability of Type II error, a statistical parameter based on the data must also be compared to a distribution that could occur if the alternative hypothesis were true; this comparison requires that H_A includes an equality. The difference between the equality relationships in H_O and H_A expresses the minimum change the experimenter is interested in detecting reliably (i.e., with a probability of $1-\beta$). In the salamander example, if the experimenter decides that it is important to detect a 10% reduction in population density, she would state the alternative hypothesis as H_A: $\mu_E \leq 0.9 \times \mu_C$.

The requirement that H_A be bounded by an equality forces us to recognize that statistical significance is not the same thing as biological significance. Given a large enough sample and enough decimal places, even the most trivial differences between means can be declared statistically significant. To avoid this outcome, the experimenter in the salamander example has clearly chosen what differences are important to detect—declines in population density of 10% or more. This difference, the distance between the equality relationships of the two hypotheses, is called the "effect size" (see Section 6.5). For differences less than the biologically significant effect size, the experimenter is willing to provisionally accept H_O when it is actually false. Having made this decision, she can design the experiment without wasting resources on the detection of biologically trivial differences.

6.3 Why Are Type I and Type II Errors Important in Adaptive Management?

6.3.1 In applied science and management, statistical inference often leads directly to decisions

Progress in basic science has traditionally depended upon a conservative approach to the acceptance of new hypotheses. This approach is used because waiting for evidence to accumulate has little disadvantage; a true scientific hypothesis will surely be discerned by the scientific community eventually after repeated iterations of the process in Figure 6.1. On the other hand, erroneously adopting a hypothesis that later proves to be false can waste time and effort. Therefore, scientists doing basic research try to minimize the probability of Type I error by keeping α small, without much explicit concern for the magnitude of β. In basic science, failure to reject the statistical hypothesis H_O does not imply the immediate acceptance of the scientific hypothesis from which it is derived, but rather a further suspension of judgement about it. A consensus to "accept" a scientific hypothesis is achieved only after repeatedly placing it at risk by testing its predictions (Platt 1964).

In contrast, applied scientists and resource managers must be concerned with both types of error. Decision-makers pressured by time constraints often must make judgements before adequate evidence is accumulated; sometimes they have only a single statistical test of the scientific hypothesis in question (Peterman 1990a). Thus, failure to reject a statistical null hypothesis may result in the *de facto* acceptance of the null scientific hypothesis. Management actions proceed as though the null hypothesis were indeed true, despite the finite probability that this conclusion might be a Type II error (Table 6.1).

6.3.2 Actions based only on either type of error can be costly

As in other resource management contexts, experiments or monitoring programs in managed forest ecosystems frequently involve tests of hypotheses about the effect of human activity on the ecosystem. The statistical tests used to evaluate these hypotheses are usually structured on null hypotheses of no effect on ecosystem variables; for example, "H_O: Water quality in streams within 100 m of a logging road is at least as good as that in watersheds without roads." With this null hypothesis, a Type I error occurs if managers conclude that water quality is reduced when in fact it is not. The costs of Type I errors usually are borne by those interested in exploiting the resource. For example, logging companies suffer financially if their roadbuilding activities are curtailed due to an incorrect rejection of H_O.

However, the costs of a Type II error may be even greater, and they often affect the ecosystem, public health, or the public purse. With a Type II error, managers would incorrectly conclude that roadbuilding does not affect water quality. The consequences would include both the effects of reduced water quality on humans and biota, and the costs of correcting the damage. Thus, in contrast to the relaxed approach to Type II error common in basic science, applied experimentation and management require attention to both types of error (Parkhurst 1990; Peterman 1990a, 1990b).

6.3.3. Decision-makers need quantitative information about probabilities of errors of inference

Management of forest ecosystems is inherently uncertain, and the costs associated with both types of errors may be high. Decision-makers must quantify uncertainties and their associated costs, including them explicitly in their plans (Morgan and Henrion 1990). Decision analysis (Raiffa 1968), an effective method for analyzing alternative management actions in the face of uncertainty, requires estimates of the probability that both null and alternative hypotheses are actually true (i.e., probabilities for the states of nature in Table 6.1). These probabilities can be calculated using Bayesian methods of statistical inference (Bergerud and Reed, this volume, Chap. 7; Peterman and Peters, this volume, Chap. 8). While it is not possible to calculate the probability that a hypothesis is true using classical statistical tests, the probability of incurring either a Type I or a Type II error can be controlled to acceptable levels through careful experimental design and statistical power analysis.

6.4 What Factors Limit Rates of Type I and Type II Error?

6.4.1 Experimenters can control Type I error rate by choosing the critical level of α

The traditional moment of truth in statistical procedures is the significance test. The P-value calculated in most familiar statistical tests indicates the probability of obtaining a test statistic at least as extreme as the one calculated from the data, if H_O were true. The significance level is a critical value of α—the maximum probability of Type I error (rejecting H_O when it is true) that the scientist is willing to tolerate. Thus, when a P-value is less than 0.05 (the usual critical value of α), the experimenter rejects the null hypothesis with the guarantee that the chance is less than 1 in 20 that a true null hypothesis has been rejected.

This guarantee about the probability of making a Type I error implicit in significance testing is valid only if the assumptions of the test are met. Otherwise, it can be difficult to determine the actual probability of Type I error. The assumptions of various significance tests are the subject matter of most practical reference books on biometrics and will not be discussed here.

The significance test analyzes only the case in which H_O is really true (the top row of Table 6.1), and thus tells only half the story. If in fact H_O is *not* true, Type II error becomes the concern (bottom row of Table 6.1). The significance test alone provides no information about its probability.

6.4.2 Experimenters can control the Type II error rate by planning for a suitable level of statistical power

While the significance test limits the rate of Type I error, it is equally important to control the probability of Type II error. To achieve this, power analysis is used to estimate β, the probability of Type II error, and its complement, statistical power ($1–\beta$), the probability of detecting a specified treatment effect when it is present (Cohen 1992).

Statistical power is a function of several variables:

- sample size, N;
- variance of the observed quantities, s^2;
- effect size (the treatment effect the experimenter wants to be able to detect); and
- α (the maximum rate of Type I error tolerated).

In general, given any four of these five variables (power, sample size, variance, effect size, and α), the fifth can be solved for.

To understand how each of these variables influences power, consider a study designed to address the question, "Is this area good breeding habitat for wood ducks?" The experimenter knows from other studies that the ducks exhibit the strongest preference for nest cavities with a diameter of 25 cm and that the standard deviation for diameter of cavities is 10.0 cm in the mixture of tree species present. He decides to sample 20 cavities and to analyze the sample mean using a *z*-test, with the statistical hypotheses H_O: $\mu = 25$ and H_A: $\mu < 25$.

This statement of the statistical hypotheses is not adequate, because statistical power analysis requires that H_A must be bounded by an equality relationship, as discussed in Section 6.2.3. In this case, the experimenter suspects that the ducks will not use cavities with a diameter less than 22 cm, so he would like to detect reliably a difference in mean diameter of 3 cm—the *biologically significant effect size.* Thus, for power analysis, the hypotheses are H_O: $\mu = 25$ and H_A: $\mu \leq 22$, with the implicit provision that the experimenter is not concerned about rejecting H_O if the mean cavity diameter lies between 22 and 25 cm.

Figure 6.2a shows the expected sampling distribution, under H_O and H_A, for estimates of mean diameter of cavity based on samples of 20 cavities. The light shaded area under the left tail of the μ_O distribution is 5% of the total area under the curve (below 21.4 cm). If the sample mean falls in this region, the null hypothesis will be rejected. If H_O is true (i.e., the true mean is indeed 25 cm), yet the sample mean falls in this tail, a Type I error will occur.

The μ_A distribution represents the sampling distribution of estimates of the mean diameter if the equality relationship of H_A is true (i.e., $\mu = 22$). In this case, observing a sample mean under 21.4 cm will lead to the correct inference (that H_O is false). Statistical power, the probability of correctly rejecting the null hypothesis, is represented by the unshaded area under the μ_A curve ($1–\beta$). The dark area under the right portion of the μ_A curve represents β, the probability of committing Type II error. If the true mean cavity size is in fact 22 cm, yet the observed sample mean is greater than 21.4 cm, the null hypothesis ($\mu = 25$) will not be rejected and a Type II error will occur. In this example, α, the probability of Type I error, is constrained to 0.05, while β, the probability of Type II error, is 0.60 and statistical power ($1–\beta$) is 0.40. Hence, the experimenter has only a 40% chance of reliably identifying inadequate breeding habitat for wood ducks.

What can the experimenter do to increase the power of this test? One possibility is to increase sample size. Figure 6.2b shows the case in which sample size has been doubled to 40. Because increasing N reduces the standard error of the mean ($s/\sqrt{N}$), the distributions under H_A and H_O become narrower. With a narrower distribution, the 5% cutoff point for hypothesis testing increases to 22.5 cm. Therefore, the dark area under the μ_A curve representing β is now reduced from 0.60 to 0.38 of the total area. In this example, therefore, doubling the sample size has increased statistical power from 0.40 to 0.62.

Increasing the effect size of interest also increases power. Suppose the experimenter believes that wood ducks will use cavities with diameters as small as 19 cm. He would like to reject H_O reliably if the mean is ≤ 19 cm, but is content to accept H_O incorrectly if the mean is between 19 and 25 cm . Figure 6.2c shows the sampling distributions for an effect size of 6 cm, twice as large as that in Figure 6.2a. In this case, the overlap between the two distributions of estimates of the mean is reduced because μ_O and μ_A are farther apart. The dark area representing β is now 0.14 of the total area, so statistical power is 0.86. The experimenter has an 86% chance of correctly identifying unsuitable wood duck habitat, more than double the power of the design in the basic scenario.

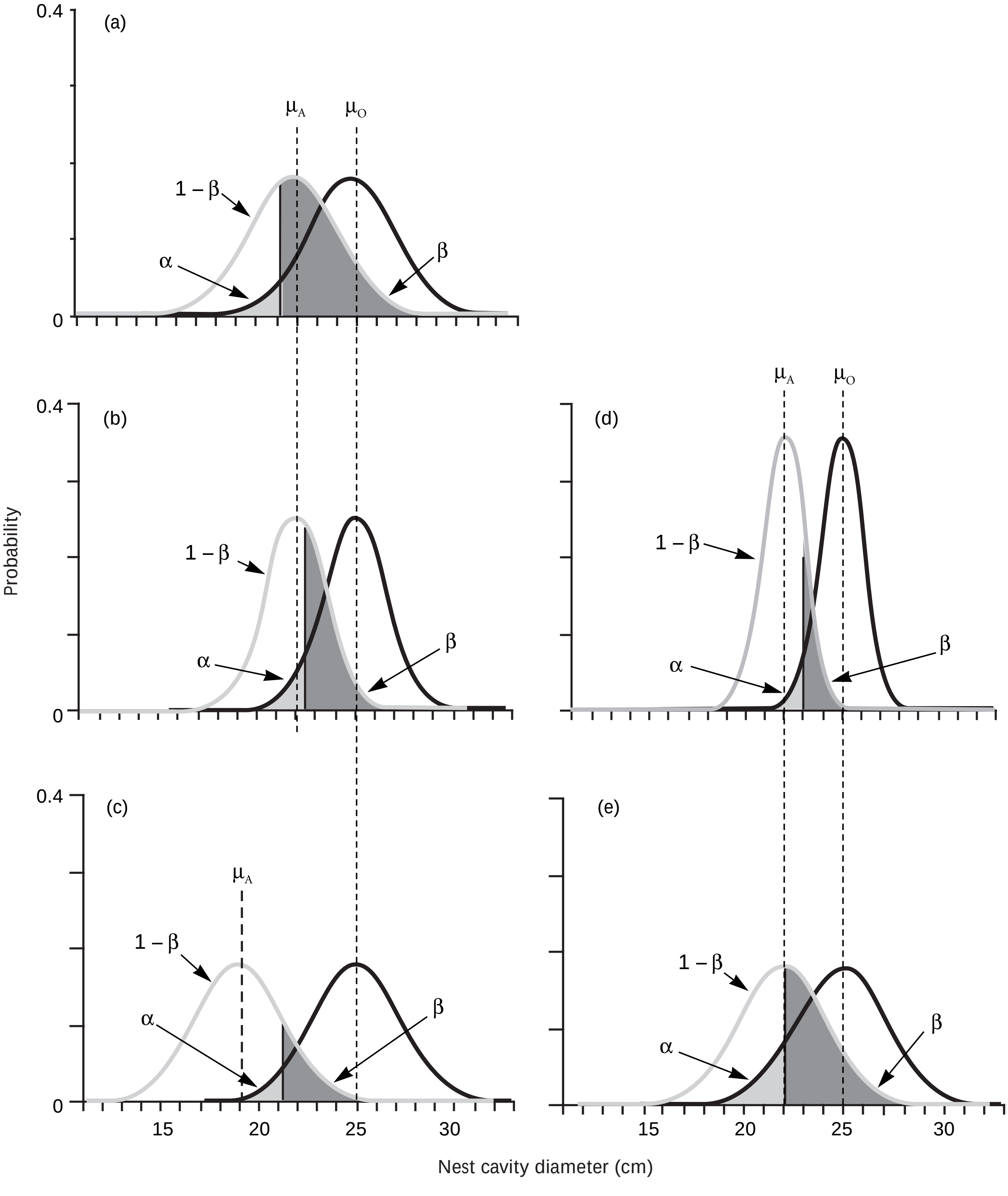


FIGURE 6.2 *Variables influencing power to detect the difference between a sample mean and a constant for the wood duck nest cavity example. For the basic scenario (a), H_O: μ=25, H_A: $\mu \leq 22$; desired effect size for detection = 3, σ = 10, α = 0.05, N = 20. Figures (b)–(e) are calculated from the basic scenario with one change in each. In (b), N = 40; in (c), effect size = 6; in (d), σ = 5; in (e), α = 0.1. The light shaded area under the H_O curve represents α, the probability of a Type I error. The dark shaded area under the H_A curve represents β, the probability of a Type II error. The unshaded area under the H_A curve represents statistical power (1–β). Each of the variations in (b)–(e) increases power relative to the basic scenario (a).*

Decreasing the variance of the observed variable (by improving measurement precision or by various techniques of experimental design, such as blocking) also increases power by reducing the standard error of the estimated means. Figure 6.2d represents the wood duck observations as in the basic scenario, with the exception that the experimenter has decided to measure only those cavities in the species of tree most favoured by the ducks. The SD is known to be 5 cm in that species, half that in the basic scenario. Here again, the dark area under the μ_A curve representing the probability of Type II error is reduced to 0.14, so statistical power is 0.86.

Finally, increasing the probability of making a Type I error will increase power. In Figure 6.2e, the critical α level (the grey tail under the μ_O distribution) is doubled to 0.10. Consequently the range of estimated means for which H_O is correctly rejected increases, resulting in a modest increase of power to 0.52. This situation requires a trade-off: the experimenter is willing to run a 10% chance of incorrectly labelling good habitats as unsuitable for wood ducks to produce a 52% chance of correctly identifying unfavourable habitats.

6.5 The Concept of Effect Size is Complex

6.5.1 How are biologically significant, detectable, and standardized effect sizes applied?

In Section 6.2.3, effect size was introduced as the difference between the equality components of the null and alternative hypotheses, usually chosen to represent a biologically significant difference. In the wood duck example, the *biologically significant effect size* of interest was easily understood as the difference between the minimum acceptable nest cavity diameter of 22 cm and the preferred size of 25 cm. This concept is most useful when the experimenter uses the relationships involving power and sample size to address a question such as, "To have an 80% chance of detecting a treatment effect at least as large as the biologically significant one, what sample size should I use?"

On the other hand, sometimes the sampling program is already set, and the experimenter would like to know how large an effect size could be detected with, say, 80% reliability. In that case, *detectable effect size* is the unknown to be calculated, rather than an input determined by biological processes or the experimenter's choice. See Section 6.6.2 for more discussion of detectable effect size.

These concepts of effect size require that all the treatment effects of the experiment be summarized in a single parameter. This approach is straightforward for some designs. In the wood duck example, the biologically significant effect size was the difference between a population mean and a known constant (the preferred cavity size). Similarly, when comparing two populations means or two correlation coefficients, the estimate of effect size is simply the difference between the two values. However, formulas for effect size become more complex in designs that involve many relationships among statistical parameters, such as multiple regression or analysis of variance.

In addition, to facilitate calculation of power and comparison between experiments, formulas for effect size are usually presented in a standardized form, including measures of variance as well as summaries of the magnitude of treatment effects. For example, the difference between two means is expressed as a *standardized effect size* by dividing by the standard deviation: $(\mu_1–\mu_2)/\sigma$, where μ_1, and μ_2 indicate the true population means and σ indicates the (common) standard deviation of the populations (Cohen 1988). Standardized effect size has several advantages. It combines into a single parameter two of the four variables that influence power—effect size and variance. Because it has no units, standardized effect size allows comparisons to be made among different experiments.

6.5.2. Arbitrary effect sizes can also be useful

Ideally the effect size to be detected should be "biologically significant," but in many cases this value cannot be expressed quantitatively due to lack of information. To assist in determining sample size in these cases, Cohen (1988) identifies representative "small," "medium," and "large" standardized effect sizes, based on the range observed in the behavioural sciences literature. Because they convey an intuitive sense about the data irrespective of the subject area, these levels can be useful as standard, if arbitrary, substitutes for a biologically significant effect size in ecological power analysis as well. In Cohen's (1988) classification system, effect sizes are typically categorized as "small" when they are subtle. Small effect sizes are often associated with newly detected phenomena or very noisy systems. "Medium" effect sizes are large enough to be perceived in the course of normal experience, while "large" effect sizes are easily perceived at a glance.

To get a feeling for small, medium, and large standardized effect sizes for the wood duck example, suppose the investigator really had no idea what deviation from the preferred mean of 25 cm would constitute an "unsuitable habitat." He might decide to detect a "medium" effect size, reasoning that a difference perceptible to a human observer should be evident to the ducks. Cohen (1988) defines a "medium" standardized effect size for a *t*-test as $(\mu_1 - \mu_2)/\sigma = 0.5$. Using the standard deviation among cavity diameters (10 cm) in the formula, the experimenter would need to set H_A: $\mu_2 \leq 20$ cm. If the experimenter felt that even a subtle difference in cavity size could be important to the ducks, he would choose Cohen's "small" standardized effect size (0.2) and set H_A: $\mu_2 \leq 23$ cm. Finally, if he felt that only an obvious difference was worth detecting reliably, he could choose a "large" standardized effect size (0.8) and set H_A: $\mu_2 \leq 17$.

Because of their importance and subtlety, effect size concepts have received considerable attention. Further discussion can be found in: Toft and Shea (1983); Rotenberry and Wiens (1985); Tanke and Bonham (1985); Stewart-Oaten, et al. (1986); Kraemer and Thiemann (1987); Millard (1987a); Cohen (1988, 1992); Forbes (1990); Parkhurst (1990); Peterman (1990a); Fairweather (1991); Faith et al. (1991); Matloff (1991); McGraw and Wong (1992); Nicholson and Fryer (1992); Schrader-Frechette and McCoy (1992); Stewart-Oaten et al. (1992); Scheiner (1993); Osenberg et al. (1994); and Mapstone (1995).

6.6 How Should Power Analysis be Used in Experimental Adaptive Management?

6.6.1 Power considerations are an intrinsic part of experimental design

The selection of experimental design is largely a question of managing the factors that influence Type II error rate: sample size, variance, effect size, and α. These variables affect statistical power in different ways. In the wood duck example, halving the sample variance and doubling the biologically significant effect size improved power more than did doubling sample size or doubling the critical level of α. Thus, while α is completely under the experimenter's control and there are good reasons to choose critical α-levels other than 0.05 (see Section 6.7), changing critical α-level is not the most effective way to gain power (Lipsey 1990). Instead, as we will discuss, efficient design usually requires specific knowledge about the experimental system, which suggests that pilot studies may sometimes be essential.

Choice of sample size and distribution of sampling effort in time and space are largely under the control of the experimenter and have been the subject of much theoretical development (Pentico 1981; O'Brien 1982; Hinds 1984; Hurlbert 1984; Andrew and Mapstone 1987; Ferraro et al. 1989; Green 1989; Krebs 1989; Kupper and Hafner 1989; Lipsey 1990; Ferraro et al. 1994; Magnussen and Boyle 1995; Mapstone 1995). For many frequently used statistical tests, the sample size necessary to achieve a given level of power is easy to look up in such references as Cohen (1988) or to calculate using software packages (Thomas 1997b; Thomas and Krebs 1997). However, in large management experiments where information is costly to gather, decisions about sample size must consider sampling costs as well.

While sample size is relatively straightforward to adjust, experimenters should not overlook other aspects of experimental design that can be equally effective at reducing rates of Type II error. First, treatments can be chosen to increase effect size detectable in an experiment. For example, in a management experiment involving effects of fertilization on green-up, it is important to know when the fertilizer should be applied to maximize its effect. Second, choosing dependent variables that are sensitive to treatments or impacts can also increase effect size (Lipsey 1990). Third, the power of an experimental design can be strongly dependent upon the shapes of response curves (Lipsey 1990; Nicholson and Fryer 1992). For example, if the relationship between the independent and dependent variable is known to be relatively steep in slope over some range of the independent variable, treatments targeted to that range would produce the greatest effect size.

Taylor and Gerodette (1993) discussed an interesting example of the choice among dependent variables, in which the power to detect a decline in a Northern Spotted Owl population was influenced by the variable monitored. Based on power analysis of a number of simulated monitoring programs, the authors made specific recommendations about the choice of dependent variable. At low population densities, estimates of birth and death rates provided higher power to detect population declines than did estimates of population size, whereas the reverse was true at high population densities.

In another study designed to identify the most useful variables for assessing impacts on water quality, Osenberg et al. (1994) compared the sensitivity of population-based biological variables, individual-based biological variables, and chemical-physical variables. They found that standardized effect sizes (and hence power) were greatest for individual-based biological variables because they responded most sensitively to impacts and exhibited relatively low error variance.

Ecological variables often exhibit large variance, so strategies for reducing variance are especially important for achieving high statistical power. The total variance is the sum of error variance and variance that can be accounted for with additional information. Therefore, it is useful to account for as much of the variance as possible in the experimental design. Examples include blocking (Hurlbert 1984; Krebs 1989), covariate analysis (Wiens and Parker 1995), and controlling for spatial heterogeneity (Dutilleul 1993). Error variance can also be reduced by improving measurement precision and reliability (Williams and Zimmerman 1989; Lipsey 1990).

How could these strategies be applied to the salamander example from Section 6.2.2? The experimenter might suspect that population density is affected by aspect, irrespective of the width of the riparian reserve. The sample watersheds could be divided into three blocks (north-facing, south-facing, and east- or west-facing), to control for the variance associated with aspect. Similarly, the experimenter could include, as covariates, information about the quantity and decay state of coarse woody debris in each sample plot to account for some of the variance. The error variance in estimates of population density might also be reduced by sampling more intensively. Finally, the experimenter might decide to estimate recruitment and death rates of salamanders in addition to population density in an effort to measure more responsive variables.

Finally, the most sophisticated design is not always the most powerful. In more complex experimental designs, statistical power is really a function of degrees of freedom, rather than straightforward sample size. Because degrees of freedom are influenced by both the extent of replication and the number of parameters to be estimated, increasing the complexity of the design can be counterproductive with respect to power. For example, in analysis of variance (ANOVA), increasing the number of factors increases the number of parameters (treatment means) to be estimated and this decreases the effective number of replicates per cell, reducing power (Cohen 1988). Therefore, to maintain comparable power to detect a given main effect within a factorial ANOVA design, more replicates per cell are required than if that same main effect were tested within a one-way design. In addition, interactions are detected with lower power than main effects. When total sampling effort is restricted, it may be impossible to increase replication adequately to justify adopting a complex design (O'Brien 1982; Cohen 1988; Lipsey 1990).

6.6.2 *A posteriori* power analysis can help interpret nonsignificant results

When interpreting a statistically nonsignificant result for a completed experiment, an experimenter should calculate power *a posteriori* rather than uncritically "accepting" the null hypothesis (Peterman 1990b; Thomas 1997a). *A posteriori* power analysis answers the question, "What was the probability that this experiment could have detected a specific, biologically important effect size?" Calculating power *a posteriori* is of course the only option if the Type II error rate was not considered in the design of the experiment, but it may also be necessary in well-designed experiments when sample variance turns out to be higher, or samples smaller, than expected. Examples of *a posteriori* power analysis are becoming increasingly common in ecological literature (Thompson and Neill 1991, 1993; Greenwood 1993; DeGraaf 1995; Reed and Blaustein 1995; Lertzman et al. 1996; Pattanavibool and Edge 1996).

Effect size can be problematic in *a posteriori* analyses. It is tempting to use the actual observed difference as a measure of effect size, but the analysis is meaningful only if based on an effect size chosen independently of the data, such as a biologically significant effect size (Thomas 1997a). If there is no obvious biologically significant effect size, Cohen's (1988) representative small, medium, or large standardized effect sizes can be used.

Alternatively, in the absence of a biologically significant effect size,non-significant results can be evaluated *a posteriori* with "reverse power analysis," which addresses the question, "What effect size could have been detected with acceptable power?" This technique involves solving for a different variable, detectable effect size rather than power (Underwood 1981). Lertzman (1992) and Schieck et al. (1995) demonstrate examples of this analysis.

Frequently, it is hard to decide what constitutes "acceptable power" in a reverse power analysis. In this case, it is more informative to present the results not as a point calculation on some arbitrarily selected level of power, but rather as a graph of power versus detectable effect size for the experiment under consideration. Figure 6.3 demonstrates this relationship for the nonsignificant correlation coefficients between density of trees growing in gaps and gap area (Lertzman 1992), based on sample sizes of 36–38. Lertzman (1992) reported that he could have detected correlations of 0.45 with 80% probability; the graph presents a wider picture (e.g., the analysis could have had a 50% chance of detecting correlations of 0.32 and a 95% chance of detecting correlations of r = 0.54 or larger).

6.6.3 Principles of sound experimentation increase power and improve inferences about causation in nonexperimental situations

When the management problem does not demand or allow direct experimentation, managers can still learn from passive (nonexperimental) monitoring, if principles of experimentation can be applied to decrease the probability of erroneous conclusions. Therefore, information-gathering protocols should be designed so that they include clearly contrasting treatments, controls, and attention to measurement error even where the situation permits only passive monitoring. The ongoing development of Before-After-Control-Impact paired (BACI) designs in environmental impact research is an instructive example of this approach. These sampling designs have been proposed to bring some benefits of experimentation to the monitoring of unreplicated environmental impacts, especially those impacts that are accidental (Bernstein and Zalinski 1983; Faith et al. 1991; Underwood 1991, 1994; Stewart-Oaten et al. 1992; Schroeter et al. 1993; Osenberg et al. 1994). See Schwarz (this volume, Chap. 3) for a discussion on sampling designs for impact studies.

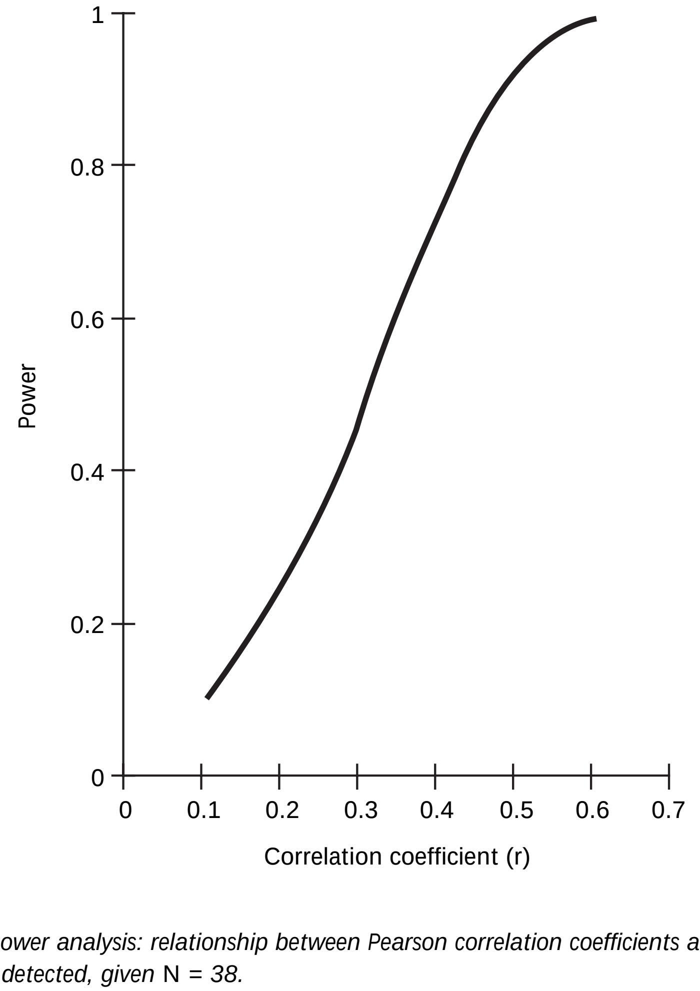


FIGURE 6.3 A posteriori *power analysis: relationship between Pearson correlation coefficients and the power with which they could be detected, given* N = *38.*

6.7 How do Experimenters Decide What Probabilities of Error are Acceptable?

6.7.1 Error rates have traditionally been set by convention

Errors of inference are inherent in any system that involves learning from incomplete information, but managers and experimenters can control the frequency of these errors. While the Type I error rate α has conventionally been limited to 0.05, no universally accepted limit for Type II error rates has emerged. Cohen (1988) suggested that experiments should be designed to have a power of 0.80. Others have proposed that α and β should be made equal (Lipsey 1990). These conventions are arbitrary and do not reflect any measure of the relative importance of the two types of error or the gravity of their consequences.

6.7.2 In applied science, costs of errors should be taken into account

In resource management experiments, where errors of inference may lead to substantial costs, those costs must be considered. Certainly the status quo, in which β is ignored, is undesirable because the probability of incurring large and possibly irreversible costs associated with Type II error can be very high and unknown.

One possibility is to set the expected costs of the two types of errors equal, where expected cost is the cost of the error times its probability (Toft and Shea 1983; Peterman 1990a). Thus, if C_1 and C_2 are the respective costs of Type I and Type II errors, we set $C_1 \alpha = C_2 \beta$. An acceptable value for β should then be $(C_1/ C_2)\ \alpha$. Mapstone (1995) suggests that both α and β may need to vary when costs of sampling or other circumstances do not allow β to become as small as prescribed by this formula. While issues of cost and error rates have been discussed by many authors (Pentico 1981; Hinds 1984; Mapstone and Andrew 1987; Millard 1987a, 1987b; Ferraro et al. 1989; Loftis et al. 1989; Davis and Hardy 1990; Holland and Ordoukhani 1990; Parkhurst 1990; Fairweather 1991; Schroeter et al. 1993; Power at al. 1995; Wiens and Parker 1995), few examples of these considerations are being practiced.

6.7.3 Other factors should also influence the choice of α and β

The purpose of the research should be considered when setting α and β. The efficiency of exploratory research or "data snooping" will often improve if α is set relatively high and β relatively low, because it is important to detect previously unknown relationships. Careful follow-up studies can discover any cases of Type I error—incorrect identification of a relationship where in reality none exists.

Another issue in setting error rates is the prior probability that each hypothesis is true. A hypothesis that seems likely to be true, based on related research, should be treated more cautiously with respect to erroneous rejection than a hypothesis that seems less credible (Lipsey 1990). For example, suppose that we can reasonably expect the community of soil microorganisms to be less disturbed in a retention silvicultural system than in large clearcuts, but that the difference may be subtle. An experiment comparing these silvicultural systems should be designed with high power to detect the difference (i.e., β should be set low).

6.8 Recommendations for Adaptive Management

6.8.1 Explicit attention to rates of Type I and Type II error should become standard practice

Several authors have surveyed experimental literature and found few examples that address Type II error (Sedlmeier and Gigerenzer 1989; Peterman 1990a, 1990b; Fairweather 1991; Searcy-Bernal 1994). Although there has been some discussion about reconsidering the arbitrary limit on Type I error ($\alpha < 0.05$), that limit is rarely reviewed in discussions of either power analysis or significance testing. However, the number of journal articles reporting new theoretical developments in the application of power analysis to ecological problems and the increasing availability of software for the purpose suggest that errors of inference will be easier to estimate and interpret in the future. The following recommendations apply to all ecological research, but especially to large-scale management experiments:

- Experimenters and decision-makers should embrace the concept that some errors of inference are unavoidable, but their frequency can be controlled by astute design of experiments and monitoring systems.
- *A priori* power analysis, with an explicit statement of desirable levels of α and β, should be included in the design process for all experiments and monitoring programs.

- All reports of nonsignificant results should mention the effect size and power of the experiment. Where appropriate, *a posteriori* power analysis may be used.

- Where potential costs of the errors of inference to various stakeholders can be quantified, these costs should be included in decisions about acceptable levels of α and β.

- Where currently available experimental designs lack power, research should be directed toward developing new, powerful methodologies, such as Before-After-Control-Impact paired designs (Underwood 1994).

- Resources should be allocated to pilot studies that will help to improve the power of large experiments.

- *A priori* power analyses are often difficult because not enough is known about potential response variables, biologically significant effect sizes, and spatial and temporal variability. It would be useful to carry out large-scale, long-term monitoring of these variables in forest ecosystems, with the express purpose of estimating them for use in future power analyses and choices about experimental design (Osenberg et al. 1994). Standard response variables such as those proposed by Keddy and Drummond (1996) for eastern deciduous forests could become a starting point for this sort of database.

6.9 Relevant Literature and Software

6.9.1 A few key references guide experimenters through power analysis for the most frequently used statistical tests

The classic reference to statistical power is Cohen (1988). Cohen provides clearly written instructions for calculating standardized effect size and other input parameters to power and sample size tables. He provides these tables for *t*-tests, tests involving correlation coefficients, tests involving proportions, the sign test, chi-square tests for goodness of fit and contingency, analysis of variance and covariance, multiple regression and correlation, and set correlation and multivariate methods (e.g., canonical correlation, MANOVA, and MANCOVA).

Zar (1996) presents a graph of power and sample size for analysis of variance, as well as formulas for calculating power and required sample size for a variety of other tests. While he does not include tabled values, the formulas are discussed with the details of the tests themselves, including biological examples. In addition, Zar discusses examples of *a posteriori* power analyses.

Nemec (1991) introduces power analysis using examples from forest research. Included are example routines for the SAS statistical software package that compare the power of completely randomized and randomized block analysis of variance designs, and calculate power for one- and two-sample *t*-tests and one- and two-way ANOVA. Several pamphlets addressing various aspects of power analysis are also available from the B.C. Ministry of Forests (Bergerud 1992, 1995a, 1995b, 1995c, 1995d; Sit 1992).

Lipsey (1990) discusses and compares many factors that affect statistical power, including adjustment of critical α and strategies for optimizing standardized effect size. This book presents power and sample size relationships as graphs, at some cost to precision in reading off the numbers.

6.9.2 Power analysis is available in many software packages

Over the last few years, the variety of software packages that perform power analysis, sample size determination, and effect size operations has greatly increased. Thomas (1997b) maintains an annotated list of software packages on the World Wide Web. Many of these packages are reviewed in Thomas and Krebs (1997). For other discussions of software, see Goldstein (1989), Borenstein et al. (1990, 1992), Rothstein et al. (1990), Steiger and Fouladi (1992), and Meyer (1995). In addition, on-line power calculations are available for ANOVA (Friendly 1996) and for correlation coefficients and tests of parameters for normal, Poisson, and exponential distributions (Bond 1996).

6.9.3 Power analysis procedures have been developed for many specialized applications relevant to ecological experimentation

Numerous examples are available of power analyses for complex designs, including factorial and repeated measures—analysis of variance (O'Brien 1982; Bittman and Carniello 1990; Muller et al. 1992), moderated multiple regression (Stone-Romero et al. 1994), and multivariate general linear hypotheses (Muller and Peterson 1984). These papers focus on practical methods for addressing questions related to power, effect size, and sample size.

Long-term ecological experiments and monitoring programs, like long-term medical trials, may not need to be carried to completion if early data are sufficiently decisive (Allison et al. 1997). These studies should therefore be subject to interim analysis to address the question, "Given the results so far, what are the chances that the conclusions will change if the experiment or monitoring is continued to completion?" As an example, Davis and Hardy (1990) described one decision process for interim analysis of long-term medical trials. This process—*stochastic curtailment*—involves calculating, given the current data, the maximum possible levels of α and β that would occur if the experiment were carried to its conclusion. If, at some interim point, those maximum error levels are suitably low, the experiment or monitoring program could be terminated.

Deviations from normal distributions are often of concern to ecologists. Gingerich (1995) compared the power of three tests designed to distinguish normal from lognormal distributions. Based on Monte Carlo simulations, he concluded that the Anderson-Darling test is the most powerful, but even it requires very large sample sizes to achieve high power. The power of all the tests was sensitive to the coefficient of variation ($V = s/\bar{x}$) in the sample; power increases as V increases. Sawilowsky and Blair (1992) discussed the Type I and Type II error properties of the *t*-test applied to measures with frequency distributions that differ grossly from normal. They concluded that the *t*-test is subject to increased Type I error under deviations from normality when the test is one-tailed or when sample sizes are small or unbalanced. Type II error does not increase dramatically, but Sawilowsky and Blair suggested that nonparametric substitutes often have higher power than the *t*-test when its assumptions of normality are seriously violated.

Measures of biodiversity are often important response variables in adaptive management experiments. To estimate statistical power for studies involving these measures, their sampling variance and other distributional properties must be known. Methods to address this issue are currently under development. For example, Magnussen and Boyle (1995) provide guidelines for power and sample size for tests involving the Shannon-Weaver and Simpson diversity indices. Mark-recapture methods are being adapted to improve the estimation of species richness and its distributional properties (E. Cooch, Simon Fraser University, pers. comm., 1996).

Other types of statistical tests, specifically designed for ecological contexts, have been subjected to power analysis. Some examples that may be useful in adaptive management experiments include:

- density dependence—power and sample size for tests designed to detect whether population parameters vary as a function of population density (Solow and Steele 1990; Dennis and Taper 1994);
- trend detection—power and sample size for tests designed to detect whether a variable is changing with time (Hinds 1984; Tanke and Bonham 1985; Harris 1986; Gerodette 1987, 1991; Whysong and Brady 1987; Kendall et al. 1992; Loftis et al. 1989);
- detection of rare species—sample sizes necessary to detect rare species, based on the Poisson distribution (Green and Young 1993);
- resource selection—patterns of Type I and Type II errors for four tests of habitat/resource selection (Alldredge and Ratti 1986);
- home range independence—power of Schoener statistic for independence of animal home ranges (Swihart and Slade 1986);
- environmental monitoring—power, sample size, and cost considerations for programs of environmental impact monitoring (Skalski and McKenzie 1982; Millard 1987b; Ferraro and Cole 1990; Ferraro et al. 1989; Parkhurst 1990; Smith and McBride 1990; Ferraro et al. 1994; Wiens and Parker 1995);
- analysis of covariance in environmental monitoring—analysis of Type I and Type II error for ANCOVA with examples from water quality monitoring (Green 1986);
- environmental impact detection, unique cases, and before-after-control-impact design issues (Bernstein and Zalinski 1983; Faith et al. 1991; Underwood 1991, 1994; Stewart-Oaten et al. 1992; Schroeter et al. 1993; Osenberg et al. 1994; Allison et al. 1997; Gorman and Allison 1997); and
- spatial patterns and heterogeneity—power analysis for experimental designs that take spatial patterns and heterogeneity into account (Andrew and Mapstone 1987; Downing and Downing 1992; Scharf and Alley 1993).

6.9.4 When analytic methods are not appropriate, Monte Carlo simulation can be used to estimate power

Many ecological analyses involve specialized statistics or experimental designs for which no analytic formulas exist for calculating power. In such cases, Monte Carlo simulation can be used to produce many simulated data sets generated from distributions with known parameter values corresponding to given null and alternative hypotheses. The experimenter can then estimate statistical power by tallying the frequency with which H_O is correctly rejected by the simulated data.

Taylor and Gerodette (1993) used this method to determine the power to detect population declines in the Northern Spotted Owl when the monitored variables were estimates of survival and reproductive rates. They simulated hypotheses about trends in population density: the null hypothesis represented a stable population ($\lambda = 1$, where λ is the geometric rate of population change), while a set of alternative hypotheses represented populations declining at specified rates (e.g., $\lambda = 0.96$). Survival, reproduction, capture, and recapture of individual birds in each declining or stable population were modelled stochastically using known ranges of variation over a number of years; the mark-recapture estimates of demographic parameters were analyzed to generate an estimate of λ. Each population and its associated estimates of λ were simulated several thousand times. For stable populations (i.e., ones for which H_O was true), the critical value of estimates of λ was determined by finding the fifth percentile of the frequency distribution of simulated estimates. Below that value, H_O would be rejected at $\alpha = 0.05$. Then, for simulated populations in which H_A was true, the power to detect the decline could be estimated by tallying the proportion of parameter estimates for which H_O was rejected at the critical value.

Other examples of Monte Carlo methods for power analysis include Loftis et al. (1989), who used it to identify the most effective tests for detecting trends in long-term water quality monitoring, and Alldredge and Ratti (1986), who simulated the statistical behaviour of a number of models of habitat selection. Using Monte Carlo simulations of fish populations, Peterman and Routledge (1983) estimated the power of proposed management experiments, and Peterman and Bradford (1987) demonstrated the difficulty of reliably detecting time trends in recruitment.

6.10 Conclusions

Scientists and resource managers are benefiting from a new level of sophistication in the statistical interpretation of management experiments and monitoring programs. This involves an explicit and balanced focus on errors of inference, including Type II error (failure to reject a false null hypothesis) as well as Type I error (rejecting a true null hypothesis). As this new standard of practice evolves, decision-makers and experimenters will find it helpful to bear in mind the following points:

- Experimenters can control statistical power *a priori* by making wise decisions regarding pilot studies, sampling methodology, and experimental or monitoring design. The advice of statistical consultants is increasingly important in the planning process as standards of experimentation improve.
- If the outcome of an experiment or monitoring program is likely to influence a management decision, statistical power should be reported when results are non-significant.
- In many situations, it is useful to free Type I and Type II error rates from traditional arbitrary values. Desirable levels of both error rates can be selected rationally as a function of the goal of the study, current understanding of the system under investigation, and costs of each type of error.
- Software and written explanations, designed to make statistical power analysis accessible to a wide range of users, are widely available.

Acknowledgements

Thanks to all those who helped with this chapter. Randall Peterman, Vera Sit, and Brenda Taylor provided much help and encouragement on the early drafts. Doug Steventon, Rita Winkler, Wendy Bergerud, Jeff Stone, Keith Thomas, Randall Peterman, and Michael Stoehr thoroughly reviewed the manuscript and provided valuable suggestions for improving the final version.

References

Alldredge, J.R. and J.T. Ratti. 1986. Comparison of some statistical techniques for analysis of resource selection. J. Wildl. Manage. 50:157–65.

Allison, D.B., J.M. Silverstein, and B.S. Gorman. 1997. Power, sample size estimation, and early stopping rules. *In* Design and analysis of single case research. R.D. Franklin, D.B. Allison, and B.S. Gorman (editors). Erlbaum, Mahwah, N.J., pp. 335–72.

Andrew, N.L. and B.D. Mapstone. 1987. Sampling and the description of spatial pattern in marine ecology. Oceanography and Marine Biology Annual Reviews 25:39–90.

Bergerud, W. 1992. A general description of hypothesis testing and power analysis. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 37.

______. 1995a. Power analysis and sample sizes for completely randomized designs with subsampling. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 49.

______. 1995b. Power analysis and sample sizes for randomized block designs with subsampling. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 50.

______. 1995c. Programs for power analysis/sample size calculations for CR and RB designs. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 51.

______. 1995d. *Post-hoc* power analysis for ANOVA *F*-tests. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 52.

Bergerud, W.A. and W.J. Reed. [n.d.] Bayesian statistical methods. This volume.

Bernstein, B.B. and J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. J. Environ. Manage. 16:35–43.

Bittman, R.M. and M.L. Carniello. 1990. The design of an experiment using statistical power with a startle chamber study as an example. J. Appl. Toxicology 10:125–8.

Bond, J. 1996. Online power calculator. Stat. Cons. Cen., UCLA, Los Angeles, Calif. Available online:</www.stat.ucla.edu/calculators/powercalc.>

Borenstein, M., J. Cohen, H.R. Rothstein, S. Pollack, and J.M. Kane. 1990. Statistical power analysis for one-way analysis of variance: a computer program. Beh. Res. Methods, Instr. Comp. 22:271–82.

______. 1992. A visual approach to statistical power analysis on the microcomputer. Beh. Res. Methods, Instr. Comp. 24:565–72.

British Columbia Ministry of Forests and B.C. Environment. 1995. Riparian area management guidebook. Victoria, B.C. Forest Practices Code guidebook.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Lawrence-Erlbaum, Hillsdale, N.J.

______. 1992. A power primer. Psychol. Bull. 112:155–9.

Davis, B.R. and R.J. Hardy. 1990. Upper bounds for Type I and Type II error rates in conditional power calculations. Commun. Statist.: Theory Methodol. 19:3571–84.

DeGraaf, R.M. 1995. Nest predation rates in managed and reserved extensive northern hardwood forests. For. Ecol. Manage. 79:227–34.

Dennis, B. and M.L. Taper. 1994. Density dependence in time series observations of natural populations: Estimation and testing. Ecol. Monogr. 64:205–24.

Downing, J.A. and W.L. Downing. 1992. Spatial aggregation, precision, and power in surveys of freshwater mussel populations. Can. J. Fish. Aquat. Sci. 49:985–91.

Dutilleul, P. 1993. Spatial heterogeneity and the design of ecological field experiments. Ecology 74 (6):1646–58.

Fairweather, P.B. 1991. Statistical power and design requirements for environmental monitoring. Austr. J. Marine Freshwater Res. 42:555–67.

Faith, D.P., C.L. Humphrey, and P.L. Dostine. 1991. Statistical power and BACI designs in biological monitoring: Comparative evaluation of measures of community dissimilarity based on benthic macroinvertebrate communities in Rockhole Mine Creek, Northern Territory, Australia. Austr. J. Marine Freshwater Res. 42:589–602.

Ferraro, S.P. and F.A. Cole. 1990. Taxonomic level and sample size sufficient for assessing pollution impacts on the Southern California Bight benthos. Marine Ecol. Prog. Ser. 67:251–62.

Ferraro, S.P., F.A. Cole, W. DeBen, and R.C. Swartz. 1989. Power-cost efficiency of eight macrobenthic sampling schemes in Puget Sound, Washington, U.S.A. Can. J. Fish. Aquat. Sci. 46:2157–65.

Ferraro, S.P., R.C. Swartz, F.A. Cole, and W. DeBen. 1994. Optimum macrobenthic sampling protocol for detecting pollution impacts in the Southern California Bight. Environ. Monitoring Assessm. 29:127–53.

Forbes, L.S. 1990. A note on statistical power. The Auk 107:438–53.

Friendly, M. 1996. Power analysis for ANOVA designs. York Univ., Toronto, Ont. Available on line: <www.math.yorku.ca/SCS/Demos/power/>[January 1996].

Gerodette, T. 1987. A power analysis for detecting trends. Ecology 68:1364–72.

______. 1991. Models for power of detecting trends: a reply to Link and Hatfield. Ecology 72:889–92.

Gingerich, P.D. 1995. Statistical power of EDF tests of normality and the sample size required to distinguish geometric-normal (lognormal) from arithmetic-normal distributions of low variability. J. Theoretical Biol. 173:125–36.

Goldstein, R. 1989. Power and sample size via MSPC-DOS computers. Am. Statist. 43:253–60.

Gorman, B.S. and D.B. Allison. 1997. Statistical alternatives for single-case research. *In* Design and analysis of single case research. R.D. Franklin, D.B. Allison, and B.S. Gorman (editors). Erlbaum, Mahwah, N.J. pp. 159–214.

Green, R.H. 1986. Some applications of linear models for analysis of contaminants in aquatic biota. *In* Statistical aspects of water quality monitoring, A.H. El-Shaarawi and R.E. Kwiatkowski (editors). Elsevier, New York, N.Y.

______. 1989. Power analysis and practical strategies for environmental monitoring. Environ. Res. 50:195–205.

Green, R.H. and R.C. Young. 1993. Sampling to detect rare species. Ecol. Applic. 3:351–6.

Greenwood, J.J.D. 1993. Statistical power. Animal Behavior 46:1011.

Harris, R.B. 1986. Reliability of trend lines obtained from variable counts. J. Wildl. Manage. 50:165–71.

Hinds, W.T. 1984. Towards monitoring of long-term trends in terrestrial ecosystems. Environ. Cons. 11:11–8.

Holland, B. and N.K. Ordoukhani. 1990. Balancing Type I and Type II error probabilities: Further comments on proof of safety vs. proof of hazard. Communications in statistics: theory and methodology 19:3557–70.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211.

Keddy, P.A. and C.G. Drummond. 1996. Ecological properties for the evaluation, management, and restoration of temperate deciduous forest ecosystems. Ecol. Applic. 6:748–62.

Kendall, K.C., L.H. Metzgar, D.A. Patterson, and B.M. Steele. 1992. Power of sign surveys to monitor population trends. Ecol. Applic. 2:422–30.

Kirk, R.E. 1982. Experimental design: procedures for the behavioral sciences. Brooks/Cole, Monterey, Calif.

Kraemer, H.C. and S. Thiemann. 1987. How many subjects? Statistical power analysis in research. Sage, London, U.K.

Krebs, C.J. 1989. Ecological methodology. Harper Collins, New York, N.Y.

Kupper, L.L. and K.B. Hafner. 1989. How appropriate are popular sample size formulas? Am. Statist. 43:101–5.

Lertzman, K.P. 1992. Patterns of gap-phase replacement in a subalpine forest. Ecology 73:657–69.

Lertzman, K.P., G.D. Sutherland, A. Inselberg, and S.C. Saunders. 1996. Canopy gaps and the landscape mosaic in a coastal temperate rainforest. Ecology 77:1254.

Lipsey, M.W. 1990. Design sensitivity: statistical power for experimental research. Sage, London, U.K.

Loftis, J.C., R.C. Ward, R.D. Phillips, and C.H. Taylor. 1989. An evaluation of trend detection techniques for use in water quality monitoring programs. Environ. Res. Lab., Off. Res. Dev., U.S. Environ. Prot. Ag., Corvallis, Oreg.

McGraw, K.O. and S.P. Wong. 1992. A common language effect size statistic. Psychol. Bull. 111:361–5.

Magnussen, S. and T.J.B. Boyle. 1995. Estimating sample size for inference about the Shannon-Weaver and the Simpson indices of species diversity. For. Ecol. Manage. 78:71–84.

Mapstone, B.D. 1995. Scalable decision rules for environmental impact studies: effect size, Type I, and Type II errors. Ecol. Applic. 5:401–10.

Matloff, N.S. 1991. Statistical hypothesis testing: Problems and alternatives. Environ. Entom. 20:1246–50.

Meyer, G.E. 1995. Power & Effect: A statistical utility for Macintosh and Windows systems. Beh. Res. Methods, Instr. Comp. 27:134–8.

Millard, S.P. 1987a. Environmental monitoring, statistics, and the law: Room for improvement. Am. Statist. 41:249–53.

______. 1987b. Proof of safety vs. proof of hazard. Biometrics 43:719–25.

Morgan, M.G. and M. Henrion. 1990. Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge Univ. Press, Cambridge, Mass.

Muller, K.E., L.M. LaVange, S.L. Ramey, and C.T. Ramey. 1992. Power calculations for general linear multivariate models including repeated measures applications. J. Am. Statist. Assoc. 87:1209–26.

Muller, K.E. and B.L. Peterson. 1984. Practical methods for computing power in testing the multivariate general linear hypothesis. Comput. Statist. Data Anal. 2:143–58.

Nemec, A.F.L. 1991. Power analysis handbook for the design and analysis of forestry trials. B.C. Min. For., For. Sc. Res. Br., Victoria, B.C. Biometrics Inf. Handb. No. 2.

Nicholson, M.D. and R.J. Fryer. 1992. The statistical power of monitoring programmes. Marine Pollut. Bull. 24:146–9.

O'Brien, R.G. 1982. Performing power sensitivity analyses on general linear model hypotheses. Proc. Statistical Computing Section, Am. Statist. Assoc.

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, K.E. Abu-Saba, and A.R. Flegal. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. Ecol. Applic. 4:16–30.

Parkhurst, D.F. 1990. Statistical hypothesis tests and statistical power in pure and applied science. *In* Acting under uncertainty: multidisciplinary conceptions. G.M. von Furstenberg (editor). Kluwer Acad. Publ, Boston, Mass.

Pattanavibool, A. and W.D. Edge. 1996. Single-tree selection silviculture affects cavity resources in mixed deciduous forests in Thailand. J. Wildl. Manage. 60:67–73.

Pentico, D.W. 1981. On the determination and use of optimal sample sizes for estimating the difference in means. Am. Statist. 35:40–2.

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

______. 1990b. The importance of reporting statistical power: The forest decline and acidic deposition example. Ecology 71:3024–7.

Peterman, R.M. and M. Bradford. 1987. Statistical power of trends in fish abundance. Can. J. Fish. Aquat. Sci. 44:1879–89.

Peterman, R.M. and R. Routledge. 1983. Experimental management of Oregon coho salmon (*Oncorhynchus kisutch*): designing for yield of information. Can. J. Fish. Aquat. Sci. 40:1212–23.

Peterman, R.M. and C. Peters. [n.d.] Decision analysis: taking uncertainties into account in forest resource management. This volume.

Platt, J.R. 1964. Strong inference. Science 146:347–52.

Power, M., G. Power, and D.G. Dixon. 1995. Detection and decision-making in environmental effects monitoring. Environ. Manage. 19:629–39.

Raiffa, H. 1968. Decision analysis: introductory lectures on choices under uncertainty. Addison-Wesley, Reading, Mass.

Reed, J.M. and A.R. Blaustein. 1995. Assessment of "nondeclining" amphibian populations using power analysis. Cons. Biol. 9:129–30.

Rotenberry, J.T. and J.A. Wiens. 1985. Statistical power analysis and community-wide patterns. Am. Nat. 125:164–8.

Rothstein, H.R., M. Borenstein, J. Cohen, and S. Pollack. 1990. Statistical power analysis for multiple regression/correlation: a computer program. Educ. Psychol. Measure. 50:819–30.

Sawilowsky, S.S. and R.C. Blair. 1992. A more realistic look at the robustness and Type II error properties of the *t*-test to departures from population normality. Psychol. Bull. 111:352–60.

Scharf, P.C. and M.M. Alley. 1993. Accounting for spatial yield variability in field experiments increases statistical power. Agron. J. 85:1254–6.

Scheiner, S.M. 1993. Introduction: theories, hypotheses, and statistics. *In* Design and analysis of ecological experiments. S.M. Scheiner and J. Gurevitch (editors). Chapman and Hall, New York, N.Y.

Schieck, J., K.P. Lertzman, B. Nyberg, and R. Page. 1995. Effects of patch size on birds in old-growth montane forests. Cons. Biol. 9:1072–84.

Schrader-Frechette, K.S. and E.D. McCoy. 1992. Statistics, costs, and rationality in ecological inference. Trends in Ecology and Evolution 7:96–9.

Schroeter, S.C., J.D. Dixon, J. Kastendiek, R.O. Smith, and J.R. Bence. 1993. Detecting the ecological effects of environmental impacts: a case study of kelp forest invertebrates. Ecol. Applic. 3:331–50.

Schwarz, C.J. [n.d.]. Studies of uncontrolled events. This volume.

Searcy-Bernal, R. 1994. Statistical power and aquacultural research. Aquaculture 127:137–368.

Sedlmeier, P. and G. Gigerenzer. 1989. Do studies of statistical power have an effect on the power of studies? Psychol. Bull. 105:309–16.

Sit, V. 1992. Power analysis and sample size determination for contingency table tests. B.C. Min. For., Res. Br., Victoria, B.C. Biometrics Inf. Pamph. No. 41.

Skalski, J.R. and D.H. McKenzie. 1982. A design for aquatic monitoring programs. J. Environ. Manage. 14:237–51.

Smith, D.G. and G.B. McBride. 1990. New Zealand's national water quality monitoring network — design and first year's operation. Water Res. Bull. 26:767–75.

Solow, A.R. and J.H. Steele. 1990. On sample size, statistical power, and the detection of density dependence. J. Animal Ecol. 59:1073–6.

Steiger, J.H. and R.T. Fouladi. 1992. R2: A computer program for interval estimation, power calculations, sample size estimation, and hypothesis testing in multiple regression. Beh. Res. Methods, Instr. Comp. 24:581–2.

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

Stewart-Oaten, A., W.W. Murdoch, and K.R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67:929–40.

Stone-Romero, E.F., G.M. Alliger, and H. Aguinis. 1994. Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. J. Management. 20:167–78.

Swihart, R.K. and N.A. Slade. 1986. The importance of statistical power when testing for independence in animal movements. Ecology 67:255–8.

Tanke, W.C. and C.D. Bonham. 1985. Use of power curves to monitor range trend. J. Range Manage. 38:428–31.

Taylor, B.L. and T. Gerodette. 1993. The uses of statistical power in conservation biology: The vaquita and northern spotted owl. Cons. Biol. 7:489–500.

Thomas, L. 1997a. Retrospective power analysis. Cons. Biol. 11:276–80.

______. 1997b. A comprehensive list of power analysis software for PCs. Math. and Comp. Sci., Univ. St. Andrews, Scotland. Available online: <www.interchg.ubc.ca/cacb/power/> [Oct. 17, 1997].

Thomas, L. and C.J. Krebs. [1997]. A review of statistical power analysis software. Bull. Ecol. Soc. Am. In press.

Thompson, C.F. and A.J. Neill. 1991. House wrens do not prefer clean nestboxes. Animal Behavior 42:1022–4.

______. 1993. Statistical power and accepting the null hypotheses. Animal Behavior 46:1012.

Toft, C.A. and P.J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. Am. Nat. 122:618–25.

Underwood, A.J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. Oceanogr. and Marine Biol. Ann. Rev. 19:513–605.

______. 1991. Beyond BACI: Experimental designs for detecting human environmental impacts on temporal variations in natural populations. Australian J. Marine and Freshwater Res. 42:569–87.

______. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecol. Applic. 4:3–15.

Whysong, G.L. and W.W. Brady. 1987. Frequency sampling and Type II errors. J. Range Manage. 40:472–4.

Wiens, J.A. and K.R. Parker. 1995. Analyzing the effects of accidental environmental impacts: approaches and assumptions. Ecol. Applic. 5:1069–83.

Williams, R.H. and D.W. Zimmerman. 1989. Statistical power analysis and reliability of measurement. J. Gen. Psychol. 116:359–69.

Zar, J.H. 1996. Biostatistical analysis (3rd ed.) Prentice Hall, Upper Saddle River, N.J.

7 BAYESIAN STATISTICAL METHODS

WENDY A. BERGERUD AND WILLIAM J. REED

Abstract

This chapter provides a brief introduction and overview of the Bayesian approach to statistical inference. The Bayesian approach is particularly suitable for adaptive management since it provides a way of incorporating prior knowledge with newly acquired knowledge and, based on a model, can be used to assign probabilities to possible states of nature. These probabilities are directly useful in decision theory applications. In contrast, the familiar and well-known frequentist methods treat each experiment or study on its own, formally ignoring other information, and thus syntheses of information must be done in an *ad hoc* fashion.

Frequentist and Bayesian statistical approaches are briefly contrasted by way of a forestry example. This example is then simplified and used to demonstrate Bayesian methods. Bayesian decision theory and model building are also described.

7.1 Introduction

In most introductory statistics courses, students encounter basic concepts of the familiar frequentist paradigm, such as sampling distributions, significance tests, P-values, and confidence intervals. These concepts, which are further developed in specialized courses on regression, design of experiments, and sampling methods, form the basis of the statistical toolkit that most graduates carry with them into the world of science and management. When faced with problems involving experiments, data, and decisions, most foresters, biologists, and forest managers will naturally reach into this toolkit. However, the familiar statistical toolkit often proves inadequate for dealing with management problems. Managers make decisions in an environment of uncertainty, where the best choice among several alternatives is unknown. They often want to know the probability that a hypothesis is true, or the degree to which it is true—information that frequentist statistics does not directly provide. Despite its limitations for management, the frequentist framework is seldom questioned by practitioners, partly because the limitations of frequentist methods are rarely discussed in introductory courses, but perhaps more importantly, because practitioners know of no alternative.

Bayesian statistics, on the other hand, is an alternative theory, which is particularly suitable for adaptive management since it provides a way of incorporating prior knowledge with newly acquired data and of assigning probabilities to possible states of nature.

The essential difference between the Bayesian and the frequentist approaches hinges on the way in which uncertainty is modelled, and on the way in which randomness is conceived. To the Bayesian, uncertainty can be represented by assigning probabilities to various possibilities. These probabilities may be obtained from previous data or from the subjective beliefs of those knowledgeable of the subject matter. In contrast, a frequentist considers that probabilities can only be meaningfully assigned to events in repeatable experiments.[1] It is claimed that the road to scientific progress lies through objectively falsifying or confirming scientific hypotheses, rather than through adjusting one's beliefs concerning their plausibility. However, adaptive management involves exactly the second kind of process, as information on the consequences of management actions becomes available as part of an ongoing process.

This chapter presents a brief overview of the Bayesian approach to statistical inference and decision-making. Bayesian ideas are also contrasted with those from the dominant frequentist paradigm. Peterman and Peters (this volume, Chap. 8) discuss in more depth the application of Bayesian statistics to management decisions.

7.2 Frequentist and Bayesian Statistics

The dominant paradigm for statistical inference is based on a frequentist notion of probability, which we will call frequentist statistics. Data generated from an experiment or study are fitted by a statistical model. This model usually includes one or more unknown quantities called parameters, which are estimated during the fitting process. The fitted model

[1] An accessible discussion of some of these points can be found in Swindel (1972), Dennis (1996), and Edwards (1996), and in the Teacher's Corner of *The American Statistician*, Vol. 51, (3), pp. 241–74 (several articles).

summarizes the data and can be used to answer questions or to calculate confidence limits for the parameters. Questions are posed as hypotheses, the most frequent being the well-known null hypothesis (that a particular parameter's unknown value is zero). By considering many hypothetical replications of the experiment, statisticians can determine the behaviour of fitted parameter values or of some function of the parameters. This behaviour is described by a frequency or sampling distribution (such as the normal, *t*-, and *F*-distributions) and is used to calculate the familiar P-values and confidence intervals.

One could argue that frequentist statistics is only really applicable to the analysis of data that arise from a procedure that is intrinsically repeatable. Such a restriction would severely limit its use. For example, it would rule out almost completely its use in time–series analysis, time being universally recognized as non-repeatable. This restriction would also rule out applying frequentist statistics in many areas of forestry, because trees generally take a long time to grow and growing conditions could change over the course of an experiment. For example, an experiment involving growing various tree species would not really be repeatable, given the dependence of growth upon weather over several decades and the possibilities of site changes. However, this objection is usually overcome by observing that replication is regarded in a hypothetical sense for the purpose of interpreting results, and that even in truly repeatable experiments, the experiment is seldom actually repeated. Rather, in both cases one contemplates a universe of possible replications for the purpose of comparing the actual observed results.

In forest management, managers would like simple answers to practical questions from sampling procedures and studies, whether experimental, observational, or a combination of the two. For example, an assessment of the probability that one or more hypotheses are true, or the probability that an estimate of a parameter is reasonably close to its unknown true value, would be useful information when formulating decisions. Insofar as frequentist statistical methods do not directly provide this information (although the results of frequentist statistical analysis are often loosely misinterpreted in this way) they may be of limited use to the forest manager. Bayesian methods on the other hand can provide precisely this sort of information.

Bayesian inference allows direct computations of the probability of a hypothesis being true, or of the probability distribution of a parameter in a statistical model. An essential difference from the frequentist approach is that probabilities are assigned to possible parameter values before the study is conducted. These probabilities can incorporate knowledge gained by the investigators from their own or others' data, their experience, and even, if so desired, their intuition. However, because it is impossible to interpret such a probability as a long-run relative frequency, a wider definition of probability known as *subjective probability* must be adopted. While this definition includes the frequentist notion of probability, it is now extended or broadened and makes meaningful such questions as, "What is the probability that the Vancouver Canucks will win the Stanley Cup in 2000?" or "What is the probability that it will rain tomorrow?" These questions have no meaning in the frequentist sense (because the Stanley Cup competition is played only once in 2000, and tomorrow is similarly unique), but certainly make sense to the person in the street, whether sports fan, gambler, or simply somebody who listens to a weather forecast.

In the Bayesian paradigm, this prior information is described by a *prior probability distribution* for the model parameters and reflects personal knowledge or beliefs before the study is conducted. The data modify this distribution to produce a *posterior probability distribution,* which describes the accumulated information. This relationship, known as *Bayes' theorem,*[2] can be written in equation form as:

$$\text{posterior} = (\text{prior}) \times (\text{likelihood of data given prior}) \times (\text{constant}) \qquad (1)$$

Although a Bayesian statistician can meaningfully make statements such as: "With posterior probability 0.95 the total volume of wood in the stand lies between 46 000 and 52 000 m^3," or that the posterior probability of a certain hypothesis being true is 0.08, making such appealing conclusions bears a price—prior probabilities must be specified and the results depend upon this specification.

Specifying the prior distribution can be challenging. The practitioner who may shudder at the thought of determining a prior distribution, should be reassured that the relative importance of the newly collected data to that of the prior distribution in

[2] An informal proof of Bayes' theorem using the example in Section 7.3 is presented in Appendix 1.

determining the posterior distribution increases with the sample size of the newly collected data. Thus, when sufficiently large amounts of data are collected, the prior distribution becomes relatively unimportant and the posterior distribution depends almost exclusively on the data via the likelihood function. Often a convenient prior distribution that reflects ignorance is the *reference prior*, which roughly speaking assumes that all possible values of the parameter (reflecting all possible hypotheses) are equally probable.

Readable discussions of the fundamentals of Bayesian statistics can be found in, for example, Wonnacott and Wonnacott (1977, Chap. 19), Kennedy (1985, Chap. 12), Berry (1996), and Ellison (1996). More complete and mathematical discussions of Bayesian methods (in rough order of increasing mathematical difficulty) can be found in, for example, Cox and Hinkley (1982), Gelman et al. (1995), and Box and Tiao (1973).

The following example will help to show the differences between the frequentist and Bayesian paradigms and how the parts of the Bayesian approach work together.

7.2.1 A silviculture example

For planning purposes, suppose that a district silviculturist is interested in knowing if a recently logged cutblock will naturally regenerate within the next 2 years. If it does regenerate then treeplanting would not be necessary and resources could be directed elsewhere. In 2 years' time, the cutblock will be sampled according to a standard protocol and the results used in a decision rule to determine if planting is necessary.

Suppose quite a bit of information is already available on other, older cutblocks, which are similar in all relevant respects and thus belong to the same stratum.[3] The silviculturist might apply this information to plan the use of available resources. This process could be done informally, by using professional judgement, or formally, by developing a statistical model to predict how likely it is that planting would be necessary. The Bayesian would develop the statistical model and call it the prior distribution for the cutblock's density in 2 years' time. Regardless of its name, the model could be used to calculate the probability that planting will be necessary at that time.

The standard frequentist approach would, at least, appear to ignore this prior information when the cutblock is sampled in 2 years' time. For instance, the protocol might be to calculate the 90% confidence limits around the density estimate of the number of stems per hectare (stems/ha) in the cutblock. If the lower confidence limit was below a minimum stocking standard (MSS) then planting would proceed. Formally, this decision rule does not use the prior distribution or any other ancillary information. The silviculturist's final decision may, of course, be based on many factors in addition to the numerical decision rule results.

In contrast, the Bayesian would explicitly include the information contained in the prior distribution by combining it with the collected data via Bayes' theorem to produce a posterior distribution for the estimate of the cutblock's density. This result could then be used to calculate a posterior probability interval (also called a credibility interval). While the fundamental concepts underlying this interval are quite different from those underlying the confidence interval, this credibility interval can be used in a similar fashion. Thus a similar decision rule could be used, namely, if the lower 90% posterior probability interval is above the MSS then conclude that the cutblock is satisfactorily restocked and does not need planting.

Both statistical approaches will sample the cutblock in 2 years, and use the data to decide if planting is necessary. Formally, this sample data will stand on its own for the frequentist while the Bayesian will use all the data—not only the newly collected data but also that collected earlier.

For this example, the Bayesian posterior probability interval may be shorter than the frequentist's confidence interval because of the inclusion of prior information (and effective increase in sample size). The mean of the posterior distribution (called the *posterior mean*) is a weighted average of the prior and observed means. If these values are not too different, then the posterior mean will be similar to that of the frequentist. On the other hand, if the observed mean is quite different from that of the prior distribution, then the posterior mean is shifted from the observed mean towards the prior mean. The amount of shift depends upon the ratio of the variance of the observed mean to that of the prior distribution. If the observed mean is well known (as indicated by a small standard error) then the shift will be small. On the other hand, if the standard error is relatively large,

[3] This sort of stratum is known as a working group within the B.C. Ministry of Forests.

then the shift could be substantial because the data have not added much information to the posterior distribution.

7.2.2 Interpreting confidence limits and P-values

This section discusses the correct interpretation of confidence intervals and P-values. Rarely do these tools directly answer the research questions under consideration, but they are often incorrectly interpreted as if they do.

To illustrate, let us consider a simple estimation problem in which a forester wishes to estimate the total volume of wood in a cutblock. From the sample data obtained, a 95% confidence interval is calculated for the total volume. Let us suppose that this confidence interval ranges from 16 000 to 24 000 m^3. The correct interpretation of this interval is that if the forester were to sample again, and again and again, each time calculating a 95% confidence interval, 95 out of every 100 calculated intervals would contain the true total volume in the long run. Or, put another way, in the long run only 5% of all 95% confidence intervals (assuming everything else is correct) would miss the mark—that is, not contain the parameter (here, the total volume) being estimated.

To the manager reading the forester's report which has a 95% confidence interval of 16 000–24 000 m^3 for the total volume, it seems much the same as saying that with probability 0.95 the total volume lies between 16 000 and 24 000 m^3. Certainly most "users" (including some statisticians) would treat the result in such a fashion. However, under the frequentist paradigm, probability is the long-run relative frequency of an event in many trials. In this example, the parameter, total volume of wood, is a fixed but unknown quantity, not a random variable. The parameter is either captured or not captured in the confidence interval. Because a relative frequency could never relate to the parameter, a researcher cannot correctly talk about probabilities concerning the parameter. Nonetheless, confidence intervals will inevitably be wrongly interpreted in this way. Most non-specialists would be hard pressed to distinguish between a statement with "95% confidence" and one with "95% probability" of being correct.

The same is true of P-values. For example, a computed P-value of 0.03 will often be misinterpreted as the probability that the null hypothesis is true. In fact, this P-value should be interpreted as the probability (in the frequentist sense) of obtaining observations as discrepant or more discrepant with the null hypothesis, than those actually observed, if in fact the null hypothesis is true. In other words, if the null hypothesis were true and the experiment were repeated many times, the results of only 3% of the experiments would be as great or greater than the discrepancy observed in the actual experiment. A mental shorthand for this is to think of the P-value as a measure of evidence against the null hypothesis (with small P corresponding to strong evidence).

While these common tools are useful, they only indirectly answer most research questions. In contrast, Bayesian methods often directly answer such questions.

7.3 The Fundamentals of Bayesian Statistics

This section will briefly describe Bayesian statistical methodology, and then use a simple numerical example to demonstrate the steps involved.

7.3.1 Components of a Bayesian analysis

To make statistical comparisons or estimations, a probabilistic model must be developed to describe how the observations are generated. As in frequentist statistics, this model will involve unknowns called parameters. While frequentist methods treat these unknowns as fixed, the Bayesian approaches these unknowns as random variables. Thus, in addition to the probability model, Bayesian inferential methods require probability distributions for the unknown parameters in that model. A Bayesian analysis requires three groups of probability distributions:

1. The first is the prior probability distribution for the unknown parameters in the data probability model. These parameters are developed before any data are collected and describe what the researcher might observe when the data are collected. This prior distribution may be developed: (1) empirically, by using previously collected data (noncontroversial but technically more demanding and known as empirical Bayes); or (2) subjectively, by carefully thinking about the situation and the science involved to develop a model that captures what the researcher might reasonably expect will be observed. When very little is known beforehand, then noninformative or reference prior distributions are used. These are convenient and give all plausible values of the unknown parameters a reasonable weighting so that the chosen prior distribution has little influence on the final outcome.

2. The second group of distributions form the probability model for the data collected from the research study. Along with the data, these distributions are used to develop the likelihood. This information is then fed into Bayes' theorem along with the prior distribution to generate the third component.
3. The posterior probability distribution is the output or outcome of a Bayesian analysis and summarizes what is known. It can be used to develop credibility intervals and probabilities for interesting values of the model parameters.

The relationship between these components of a Bayesian analysis is shown in Figure 7.1.

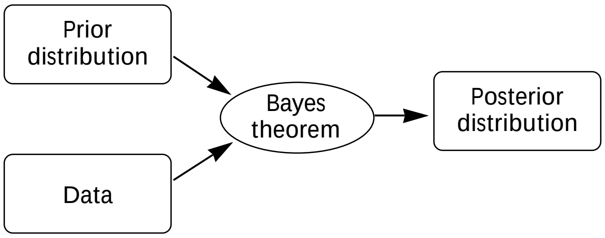


FIGURE 7.1 *Components of a Bayesian analysis.*

7.3.2 Example with single sample plot

For illustration purposes, let us rework the previous silviculture example. Recall that the silviculturist is interested in determining whether the cutblock will be satisfactorily restocked (*SR*) or not satisfactorily restocked (*NSR*). This measurement will determine if planting will be necessary. Suppose that this particular stratum has 100 previously studied cutblocks; of those, 16 were *SR* (16%) while 84 were *NSR* (84%).[4] This information provides the prior probability distribution,[5] namely the probability that the new cutblock is *NSR* is 0.84. Thus, without sampling the new cutblock, the odds are 84 to 16 or, about 5 to 1, that it will need planting.

Suppose that we put just one plot into this cutblock, and determine that it is understocked (i.e., the observed number of well-spaced trees in that plot is less than the MSS). These data are used with the probability model to develop the likelihood. We can use more of the previously collected data on the similar 100 cutblocks to build this probability model and so eventually develop the posterior distribution.

TABLE 7.1 *Numbers of previously sampled plots (observed as* US *or* S*) from both* NSR *and* SR *cutblocks. An average of 10 plots was placed in each cutblock.*

	Stocking status of plots		
Cutblock status	*US*	*S*	Total
NSR	672	168	840
SR	72	88	160
Total	744	256	1000

Suppose that an average of 10 plots was placed in each cutblock and that each plot was classified as understocked (denoted by *US*) or stocked (denoted by *S*), with the results shown in Table 7.1. These data are also shown pictorially in Figure 7.2. For this simple

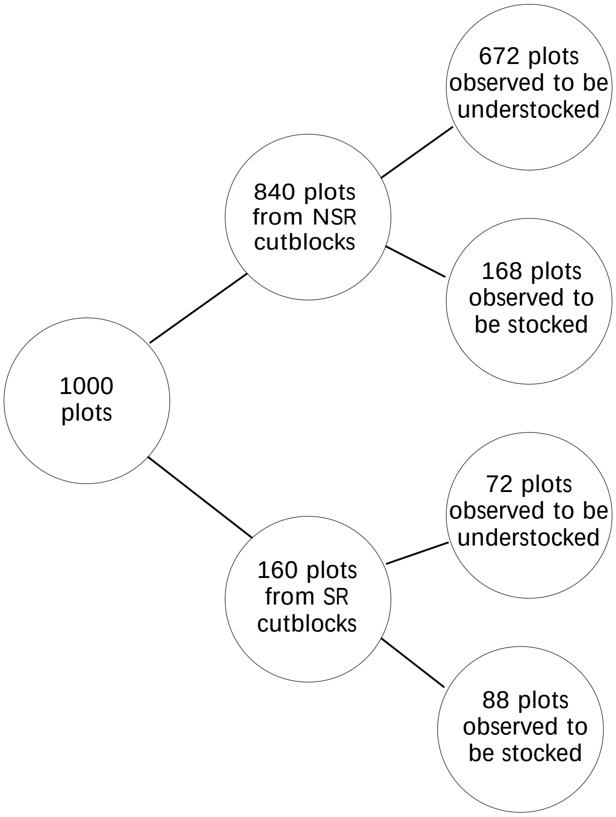

1000 plots
840 plots from NSR cutblocks
160 plots from SR cutblocks
672 plots observed to be understocked
168 plots observed to be stocked
72 plots observed to be understocked
88 plots observed to be stocked

FIGURE 7.2 *Distribution of sample plots for the silviculture example.*

[4] Assume that this information is known without error.

[5] Assuming, of course, that this new cutblock is correctly identified as belonging to this stratum.

example, we can actually skip the development of the likelihood and directly obtain the posterior distribution from Table 7.1. Notice that 744 of all the observed plots were found to be *US*. Of those, 672 were sampled from *NSR* cutblocks. Therefore, the posterior probability that the current cutblock is *NSR* given that the one observed plot is *US* is

$$\frac{672}{744} = 0.903.$$

With relevant prior information and the sampling results of just one plot, the Bayesian approach allows the determination of the posterior probability that the cutblock is *NSR*. On the other hand, the frequentist approach formally ignores the prior information and so could do little with just one plot. In any case, regardless of approach, it is unwise to decide the management of the cutblock on the basis of one plot. In Section 7.3.1, we will extend the methodology to samples of several plots. For the rest of this section, we will fill in the steps just skipped by developing the necessary statistical notation, the probability model for the data, and the posterior probability distribution.

The two possible "true" states of the cutblock can be denoted by the random variable θ, which can have one of two values, either *NSR* or *SR*. Based on Table 7.1, the prior probability that the cutblock is *NSR* is denoted by $p(\theta = NSR) = \pi_o = 0.84$ while the prior probability that the cutblock is *SR* is denoted by $p(\theta = SR) = 1–\pi_o = 0.16$. This prior distribution has a Bernoulli distribution with parameter π_o (a special case of the binomial distribution when the sample size is one).

The data, denoted by *X*, can have one of two values: either $X = US$ or $X = S$. The probability model provides the probability of the data given the "true" state of nature and is denoted by $p(X|\theta)$ (the vertical bar "|" means "given"). This symbol is read as "the probability of observing the data, *X*, given (a specific value of) the parameter, θ." In this case, two such models are used—one for each state of nature, θ. For a sample of one plot, both models follow a Bernoulli distribution. From Table 7.1, we see that

$$p(X = US|\theta = NSR) = \frac{672}{840} = 0.80\ ,$$

while

$$p(X = US|\theta = SR) = \frac{72}{160} = 0.45.$$

These probabilities are parameters of the probability models for the two states of nature and will be denoted by π_1 and π_2, respectively. They are also *conditional probabilities* because they are the probability that $X = US$ given (or conditional on) the true state of nature. Furthermore, these probabilities are the *likelihood*[6] functions, because they provide a measure of the likelihood of observing the data, *X*, given specific values for the state of nature, θ. The model parameters are summarized in Table 7.2 and in the tree diagram in Figure 7.3.

A derivation of Bayes' theorem using this example is presented in Appendix 1. The conclusion from that appendix is that the posterior probability of a particular state of nature, θ, given the data *X* is given by:

$$p(\theta/X) = \frac{p(\theta) \times p(X/\theta)}{p(X)}.$$

For this example, the probability that the new cutblock is *NSR* given that the observed plot is *US* is:

Table 7.2 The probability parameters for $p(\theta)$*, the prior distribution of* θ *(cutblock is* NSR *or* SR*), and* $p(X|\theta)$*, the conditional probability of observing* X*, given* θ *(cutblock is* NSR *or* SR*). All values were calculated from those in Table 7.1.*

Probability parameters		Conditional probabililty or likelihood, $p(X/\theta)$ for	
Cutblock status	Prior probability: $p(\theta)$	$X = US$	$X = S$
$\theta = NSR$	$\pi_0 = 0.84$	$\pi_1 = 0.80$	$1–\pi_1 = 0.20$
$\theta = SR$	$1–\pi_0 = 0.16$	$\pi_2 = 0.45$	$1–\pi_2 = 0.55$

[6] Not all conditional probabilities are likelihoods.

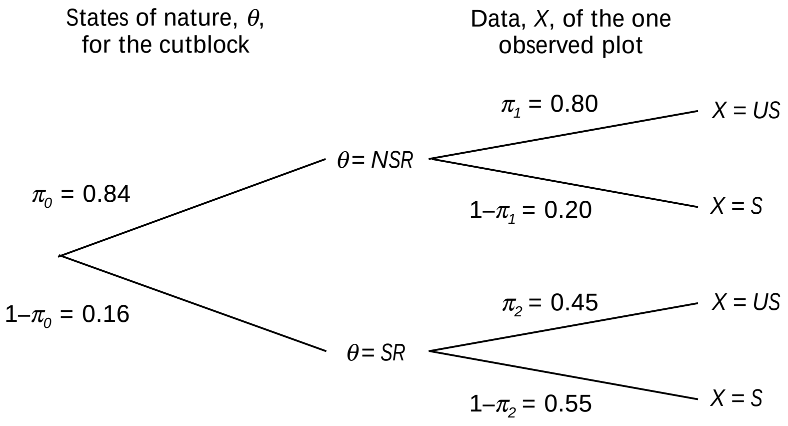


FIGURE 7.3 *Probability tree for the silviculture example.*

TABLE 7.3 *The likelihoods (probability that* X *plots out of 12 are* US *given* θ*) and the posterior probability that the cutblock is* NSR *for all possible* X *values, when the prior probability,* π_0 *= 0.84*

X = number of US plots observed	Likelihood that plots are US when $\theta = NSR$ p($X \mid \theta = NSR$)	Likelihood that plots are US when $\theta = SR$ $p(X \mid \theta = SR)$	Posterior probability for $\theta = NSR$ $p(\theta = NSR \mid X)$	Cutblock is most likely	Management decision
0	0.000	0.001	0.000	SR	not plant
1	0.000	0.008	0.000	SR	not plant
2	0.000	0.034	0.001	SR	not plant
3	0.000	0.092	0.003	SR	not plant
4	0.001	0.170	0.016	SR	not plant
5	0.003	0.222	0.073	unclear	unclear
6	0.016	0.212	0.277	unclear	unclear
7	0.053	0.149	0.652	unclear	unclear
8	0.133	0.076	0.902	unclear	unclear
9	0.236	0.028	0.978	NSR	plant
10	0.283	0.007	0.995	NSR	plant
11	0.206	0.001	0.999	NSR	plant
12	0.069	0.000	1.000	NSR	plant
Total	1.000	1.000			

$$\text{posterior probability} = p(\theta = NSR|X = US) = \frac{\pi_o \times \pi_1}{\pi_o \times \pi_1 + (1-\pi_o) \times \pi_2} = 0.903$$

as earlier found directly from Table 7.1.

7.3.3 Example with several sampling plots

Making management plans by sampling one plot is unrealistic. Therefore, we extend the Bayesian methods for one plot to that of several plots so that the cutblock can be properly sampled. The mathematical development for this case is given in Appendix 2. Suppose a sample of *n* plots is placed in the cutblock with the same prior probabilities as shown in Table 7.1. We will now use *X* to represent the number of plots observed to be understocked (*US*). From Appendix 2, the posterior probability that the cutblock is *NSR* given that *X* plots are observed to be *US* is:

$$p(\theta = NSR|X) = \frac{p(\theta = NSR) \times p(X|\theta = NSR)}{p(X)} . \qquad (2)$$

Table 7.3 gives the likelihoods and posterior probabilities that the cutblock is *NSR* for $n = 12$ sample plots (using a protocol of one plot per hectare of cutblock to determine that 12 plots are required) and for all possible *X* values ($X = 0$ to $X = 12$). The posterior probabilities are calculated using equation (2), with the prior probability being $p(\theta = NSR) = \pi_o = 0.84$.

If four or fewer plots out of 12 are observed to be *US*, then deciding that the cutblock is *SR* seems clear because the posterior probability is less than 0.05. Also, if eight or more plots are observed to be *US* then the *NSR* decision is clear because the posterior probability is greater than 0.95. But when 5, 6, or 7 out of 12 plots are observed to be *US* then the management decision is not clear. These results depend on the model probability values of π_1 and π_2. If their values had been more widely separated then the unclear decision would have occurred for fewer *X*-values; if their values had been more similar, then the undecided decision would have occurred for more *X*-values. Odds ratios provide another way to express these results.

7.3.4 Odds ratios

Another way to look at these results is to calculate the *posterior odds* that the cutblock is *NSR* ($\theta = NSR$) given *X* plots observed to be understocked. This calculation determines the ratio of the posterior probability that the cutblock is *NSR* (denoted by $p(\theta = NSR|X)$) to the posterior probability that the cutblock is *SR* (denoted by $p(\theta = SR|X)$), namely:

$$\frac{p(\theta = NSR|X)}{p(\theta = SR|X)} = \frac{p(\theta = NSR) \times p(X|\theta = NSR)}{p(\theta = SR) \times p(X|\theta = SR)} = \frac{p(\theta = NSR)}{p(\theta = SR)} \times \frac{p(X|\theta = NSR)}{p(X|\theta = SR)} = \frac{\pi_o}{1-\pi_o} \times \frac{p(X|\theta = NSR)}{p(X|\theta = SR)} . \qquad (3)$$

The posterior odds is composed of two parts:

1. the ratio of the prior probabilities, or the prior odds is
   $$\frac{\pi_o}{1-\pi_o} = \frac{0.84}{0.16} = 5.25, \text{ and}$$
2. the ratio of the two conditional probabilities known as the *Bayes Factor*,
   $$\frac{p(X|\theta = NSR)}{p(X|\theta = SR)}$$

For $n = 1$ and $X = US$, the Bayes factor is:

$$\text{Bayes factor} = BF = \frac{p(X = US|\theta = NSR)}{p(X = US|\theta = SR)} = \left(\frac{\pi_1}{\pi_2}\right) = \frac{0.80}{0.45} = 1.78$$

and so the posterior odds are:

$$\frac{p(\theta = NSR|X)}{p(\theta = SR|X)} = \left(\frac{\pi_o}{(1-\pi_o)}\right) \times \left(\frac{\pi_1}{\pi_2}\right) = \left(\frac{0.84}{0.16}\right) \times \left(\frac{0.80}{0.45}\right) = \frac{0.672}{0.072} = 9.3.$$

This value means that, having observed that one plot is *US*, the odds are about 9 to 1 that the cutblock is *NSR*. If the observed plot had been *S* (stocked), the odds would have been

$$\left(\frac{0.84}{0.16}\right) \times \left(\frac{0.20}{0.55}\right) = (5.25 \times 0.364) = 1.91,$$

about 2 to 1 that the cutblock is *NSR*, or about 1 to 2 that the cutblock is *SR*. Notice how the data have modified the prior odds of about 5 to 1.

For multiple sample plots, $n = 12$, with seven plots ($X = 7$) found to be *NSR*, the prior odds remain at

5.25 while the Bayes factor will be (values from Table 7.3) (0.053/0.149) = 0.36. Thus the posterior odds is now (5.25 × 0.36) = 1.78, meaning that the odds are about 9 to 5 that the cutblock is *NSR*.

A value of the Bayes factor greater than 1 indicates that the data contain evidence to favour the hypothesis that the cutblock is *NSR*, while a value less than 1 indicates the opposite (evidence favours the hypothesis that the cutblock is *SR*).

A common *reference prior* (used, for instance, if no data had been available to develop the prior probabilities) would give both possible states, *NSR* or *SR*, equal probabilities so that the prior odds would be 1. In this case, the Bayes factor would directly provide the posterior odds. For further discussion, see Kass and Raftery (1995).

7.3.5 Bayes factor for significance testing

The Bayes factor is a measure of the evidence for two competing hypotheses, and can be directly used to help choose one hypothesis over the other. Example cutoff values for both are shown in Table 7.4 and were taken from Ellison (1996), who quotes Kass and Raftery (1995, Sec. 3.2).

TABLE 7.4 *Suggested cutoff values for the Bayes factor (*BF*) when comparing two hypotheses*

BF	Evidence against H: Cutblock is *SR* as opposed to *NSR*
0 – 1	Nothing to mention
1 – 3	Not worth more than a bare mention
3 – 20	Positive
20 – 150	Strong
> 150	Very strong

For the multiple sampling plot example in the last section, the Bayes factor was 0.36, which suggests little evidence against the hypothesis that the cutblock is *SR*. Alternatively, the Bayes factor for the hypothesis that the cutblock is *NSR* is 1/0.36 = 2.78, which, while still low, suggests that the evidence is more supportive of the hypothesis that the cutblock is *SR* than *NSR*. The Bayes factor can be used similarly to the P-values in hypothesis testing. An advantage of Bayes factor is that it is not sensitive to sample sizes, whereas frequentist P-values can be dramatically affected by unusually large or small sample sizes (Cox and Hinkley 1974, Table 10.2; Ellison 1996).

7.4 Bayesian Decision Theory

Both the inferential problems of estimation and hypothesis testing can be viewed as problems in decision theory, for which a complete Bayesian theory has been developed. However, Bayesian decision theory can also be used in applied problems of decision-making when information is obtained through experience and experimentation. For instance, the natural regeneration example previously discussed could be formulated as a Bayesian decision theory problem, as could many other questions relating to forest management.

The basic framework of decision theory assumes a set of possible, but unknown, "states of nature," $\Theta = \{\theta_1, \theta_2, \ldots\}$, and a set of possible actions $A = \{a_1, a_2, \ldots\}$ available to the decision-maker.[7] If the decision-maker chooses an action, a_1, when the state of nature is θ_1 then an incurred loss can be calculated by a function denoted by $L(\theta_1, a_1)$. This loss could also be written as a gain $G(\theta_1, a_1) = - L(\theta_1, a_1)$. For the natural regeneration example, the set has two states of nature: $\Theta = \{\theta_1 = NSR, \theta_2 = SR\}$. The two possible actions under consideration are $A = \{a_1 =$ plant, $a_2 =$ not plant$\}$. For illustration purposes,[8] some arbitrary numbers will be used for the gain function and are presented in Table 7.5 and Figure 7.4. This figure shows a simple decision-tree diagram often used in decision analysis. This example will be used to illustrate the basic concepts in decision analysis, which are developed in more detail by Peterman and Peters (this volume, Chap. 8).

The decision-maker wants to keep losses as small as possible or, alternately, the gains as high as possible. The difficulty lies in the fact that there is usually not a unique action, a^*, for which the gain is maximized for all states of nature, θ. For some states of nature one action maximizes the gain, while for others a different action will provide a maximum. In such cases, since the state of nature is unknown, an unambiguously "best" action cannot be chosen. For example, planting a site when it is sufficiently regenerated is a waste of resources and may require further resources later, if, for instance, the stand is too dense

[7] Θ and A are names used to represent sets of things, which consist of the possible states of nature: θ_1, θ_2,..., and the possible actions a_1, a_2 ..., respectively.

[8] Although we have used the gain function when writing this section because of its more positive point of view, the literature mostly uses the loss function.

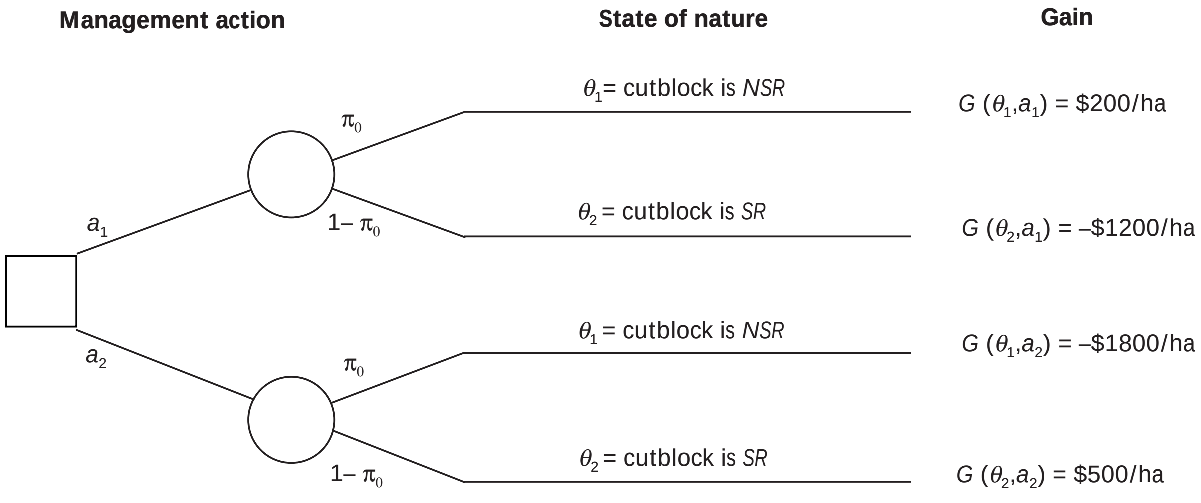

Management action
State of nature
Gain
θ_1= cutblock is *NSR*
$G(\theta_1,a_1)$ = $200/ha
π_0
a_1
$1-\pi_0$
θ_2 = cutblock is *SR*
$G(\theta_2,a_1)$ = –$1200/ha
θ_1 = cutblock is *NSR*
$G(\theta_1,a_2)$ = –$1800/ha
a_2
π_0
$1-\pi_0$
θ_2 = cutblock is *SR*
$G(\theta_2,a_2)$ = $500/ha

FIGURE 7.4 *Decision tree for the silviculture example.*

and thinning is required. On the other hand, not planting the site when needed may mean that, at rotation (harvest), the stand produces much less volume than it could have, resulting in a significant loss in revenue.

The best action, called the *Bayes action*, minimizes the expected loss, or *Bayes loss*, over all possible actions, *a*, with respect to the prior distribution. This action is equivalent to maximizing the expected gain. Table 7.5 shows some hypothetical gains for each action under each state of nature. Using the prior probability, $p(\theta)$, to model the probability of a particular state of nature, we can calculate the expected gains (or losses) for each combination of θ and *a*. The Bayes gain (*BG*) for the first action (a_1 = plant) can be calculated by:

$$\begin{aligned}BG(a_1) &= \pi_o G(\theta_1,a_1) + (1-\pi_o)G(\theta_2, a_1)\\ &= 0.84 \times \$200 + 0.16 \times (\text{-}\$1200) = \text{-}\$24/\text{ha},\end{aligned}$$

and for the second action (a_2 = not plant):

$$\begin{aligned}BG(a_2) &= \pi_o G(\theta_1,a_2) + (1-\pi_o)G(\theta_2, a_2)\\ &= 0.84 \times (\text{-}\$1800) + 0.16 \times \$500 = \text{-}\$1432/\text{ha}.\end{aligned}$$

Since the Bayes gain is greatest (-$24/ha) for the action a1 (plant), then the recommended Bayes action is to plant the cutblock.

So far, the decision has been based on the prior distribution. We can use data to update the Bayes gain to obtain a Bayes posterior gain. We can maximize this gain to choose the action, $a^*(X)$, from all the possible actions *A* calculated for every possible value of the data. Thus we would have an optimal decision rule (or policy) prescribing the optimal action for any observed data value. This policy is known as the Bayes decision rule, and can be shown to minimize what is known as the Bayes risk over all decision rules that assign an action to every possible value of the data, *X*.

Continuing the example, the Bayes posterior gain can be calculated using the posterior probability, $p(\theta = NSR|X)$, whose values are presented in Table 7.6.

TABLE 7.5 *Hypothetical gains for each combination of action and state of nature*

	Possible action	
State of nature	a_1 = plant	a_2 = not plant
$\theta_1 = NSR$	$G(\theta_1,a_1)$ = $200/ha	$G(\theta_1,a_2)$ = -$1800/ha
$\theta_2 = SR$	$G(\theta_2,a_1)$ = -$1200/ha	$G(\theta_2,a_2)$ = $500/ha

TABLE 7.6 *The Bayes posterior gain and posterior Bayes decision for all possible numbers of understocked plots*

Number of understocked plots observed	Posterior probability for $\theta = NSR$ ($\pi_p = p(\theta = NSR \vert X)$)	Posterior gain for action: a_1 = plant	Bayes posterior gain for action: a_2 = not plant	Posterior Bayes decision	Prior Bayes decision
0	0.000	-1200	500	not plant	plant
1	0.000	-1200	500	not plant	plant
2	0.001	-1199	498	not plant	plant
3	0.003	-1195	492	not plant	plant
4	0.016	-1178	464	not plant	plant
5	0.073	-1098	333	not plant	plant
6	0.277	-812	-137	not plant	plant
7	0.652	-287	-1000	plant	plant
8	0.902	62	-1574	plant	plant
9	0.978	169	-1750	plant	plant
10	0.995	194	-1790	plant	plant
11	0.999	199	-1798	plant	plant
12	1.000	200	-1800	plant	plant

Because this notation is cumbersome, we will use $\pi_p = p(\theta = NSR|X)$ for the rest of the section. For the observed data of 7 *US* plots out of 12, the Bayes posterior gain for the first action (a_1 = plant) can be calculated by

$$\begin{aligned} BG(a_1) &= \pi_p G(\theta_1, a_1) + (1-\pi_p)G(\theta_2, a_1) \\ &= 0.652 \text{ x } \$200 + 0.348 \text{ x } (-\$1200) = -\$287/\text{ha}, \end{aligned}$$

and for the second action (a_2 = not plant):

$$\begin{aligned} BG(a_2) &= \pi_p G(\theta_1, a_2) + (1-\pi_p)G(\theta_2, a_2) \\ &= 0.652 \times (-\$1800) + 0.348 \times \$500 \\ &= -\$1000/\text{ha}. \end{aligned}$$

Given this data, the best action would be to plant the cutblock if 7 out of 12 plots were observed to be *US*. Bayes posterior gains have been calculated for each value of *X* and the resulting Bayes decisions presented in Table 7.6.

The action with the higher Bayes posterior gain would be optimal, that is, it would be optimal to plant (a_1) if

$$\pi_p \times G(\theta_1,a_1) + (1-\pi_p) \times G(\theta_2,a_1) > \pi_p \times G(\theta_1,a_2) + (1-\pi_p) \times G(\theta_2,a_2),$$

or after rearranging:

$$\frac{\pi_p}{1-\pi_p} = \frac{\pi_o}{1-\pi_o} \times \frac{P(X|\theta_1)}{P(X|\theta_2)} > \frac{G(\theta_2, a_2) - G(\theta_2,a_1)}{G(\theta_1, a_1) - G(\theta_1, a_2).} \qquad (4)$$

The left-hand side is the posterior odds (see equation (2)). If it is greater than the ratio of gain differences on the right-hand side, then planting will be the Bayes decision. If this odds is less, then not planting would be the Bayes decision. Thus the condition (equation (4)) can be expressed as: plant if and only if the evidence for an *NSR* cutblock is sufficiently high. How high it has to be depends on the prior odds, and on the anticipated gains under all scenarios (via the right-hand side of equation(4)).

For our example, the ratio of gains is:

$$\frac{G(\theta_2, a_2) - G(\theta_2, a_1)}{G(\theta_1, a_1) - G(\theta_1, a_2)} = \frac{500 - (-1200)}{200 - (-1900)} = \frac{1700}{2000} = 0.85.$$

When the posterior odds are greater than 0.85 then the decision is to plant. For the example, this occurs for all *X* greater than 6. If the posterior odds is less than 0.85 then the decision is to not plant. Note that the posterior decision depends on the data while the prior decision does not (because the prior odds was 5.25 > 0.85, the prior decision was to plant). Recall that the prior Bayes decision was calculated previous to any data collection and thus is constant for all possible data values.

This basic framework can be extended in many ways. For example, in a sequential decision problem, the decision-maker can decide at each step, either to: (1) collect more data and defer choosing an action *a* from *A*, or (2) stop data collection and choose an action *a* from *A*. Sequentially at each stage in the sampling, the decision-maker can make a choice based on current data, or decide to defer choice and collect more data. Details of this and other problems in Bayesian decision theory can be found in Berger (1985). More discussions on decision analysis are presented in Peterman and Peters (this volume, Chap. 8).

7.5 Bayesian Model Building and Prediction

Modelling is a common tool for simulating the underlying processes of a system. In forestry, for example, models are developed to simulate tree growth or timber and then predict tree volume in the forests and timber supply. These predictions could be one of the factors considered in setting forest management policy, so the reliability of these models is very important. The development of models involves statistical analysis to decide which factors are important, to choose how these factors should be represented, and to validate the output of a model against observed behaviour.

In classical statistics, given a certain set of data (e.g., for each experimental unit there is a response y and a set of regressor variables $x_1, x_2, \ldots, x_p$) the first step is usually to identify a plausible model, and then use that model to answer the questions of interest to the experimenter. For example, in a forestry study y might be the volume of timber at a certain age, with the x variables corresponding to species type, spacing treatment, fertilization treatment, site index, site altitude, and site aspect for various test plots. First, some variable selection technique would be used to decide which regressor variables, with what transformations and what interaction terms, should be included in the model. After a model had been satisfactorily identified, the analyst would address such questions as the efficacy of spacing and fertilization treatments.

However, a possible weakness with this approach is that the final inferences are all contingent on the model selected. Several different models may all have a similar degree of plausibility, which could yield somewhat different predicted outcomes and lead to different decisions about treatments. Which model should you choose? This situation can be handled quite easily in the Bayesian framework. Essentially, prior probabilities are assigned to possible models, and via Bayes' theorem the data are used to obtain posterior probabilities. Then many possible models are usually eliminated by restricting attention to a few with relatively high posterior probabilities. Subsequent analysis can be carried out by averaging (with posterior probability as weights) over the set of plausible models. Thus the estimated effect of a certain silvicultural treatment would be obtained as a weighted average of the estimated effect for a number of different plausible models. Similarly, a predicted volume for a certain combination of treatments and site characteristics could be obtained using the posterior distributions of regression coefficients in each plausible model, and averaging with the posterior weights for the various models. That is, instead of using a single model, prediction would be based on a number of highly probable models. With adaptive management, managers can design management actions as experiments to distinguish between plausible models, thus improving future predictions, management decisions, and outcomes.

When the number of regressor variables is large, numerous subset models may be generated, possibly too many to handle even with modern computing power. A number of methods have been proposed to reject implausible models (described in Raftery 1994).

7.6 Conclusion

Bayesian methods provide an attractive alternative to the frequentist methods of classical statistics. The Bayesian approach systematically includes prior information in the analysis, thus better matching the manner in which we learn. Another attraction is that it permits direct computation of the probability of an hypothesis being true, and the probability that an estimate of a parameter is reasonably close to the unknown true value, hence aiding managers in decision-making. Bayesian methods also allow a common-sense interpretation of statistical conclusions, instead of the rather convoluted frequentist explanations of confidence levels and P-values. In recent years in applied statistics, interval estimation has increasingly been emphasized over the use of hypothesis tests. This shift provides a strong impetus for using Bayesian methods, because it seems highly unlikely that most users give confidence intervals anything other than a common-sense Bayesian interpretation. Furthermore, where learning and experimentation take place sequentially—as occurs in adaptive management—the Bayesian approach seems the natural way to update knowledge.

The basic steps in a Bayesian analysis are:

1. Setting up a full probability model—a distribution of all observable quantities conditional on the parameters (unobservable quantities). The extra specification that the Bayesian requires over the frequentist is a prior distribution for the parameters of the data probability model. The frequentist regards these parameters as simply unknown quantities, whereas the Bayesian regards them as random variables, and uses a probability distribution to reflect the current state of knowledge concerning their value.
2. Obtaining a posterior distribution of the parameters by conditioning on the observed data (via Bayes' theorem). In other words, obtaining the conditional probability distribution of the unobserved parameters, given the observed data.
3. While not discussed in this chapter, the fit of the model can be evaluated by answering questions such as: Does the model fit the data? and how do the conclusions depend on the modelling assumptions in step 1? If necessary the model can be revised, and the three steps repeated.

One of the strongest objections to Bayesian statistics is the requirement for a prior distribution. However, with a sufficiently large amount of data, the prior distribution becomes unimportant and the posterior probability depends almost entirely on the data. When data are scarce, all results, whether obtained by the frequentist or Bayesian methods, should be interpreted with caution.

One of the central features of the Bayesian approach is that it permits a direct quantification of uncertainty. This means that there are no impediments to fitting models with many parameters and complicated probability specifications, except for the practical ones of computing complicated multidimensional posterior distributions. However, recent advances in computing power have greatly expanded the possibilities in this area, leading to a remarkable renaissance in Bayesian statistics. The recent book by Gelman et al. (1995) provides an up-to-date exposition of the theoretical and practical aspects of modern Bayesian methodology.

Forest managers must make sound management decisions based on their knowledge of the system being managed (the system may include the forest ecosystem as well as economic and social elements) and existing data. Bayesian methods provide a way of explicitly integrating a manager's accumulated knowledge with experimental data in a statistical analysis or decision-making process.

Acknowledgements

We would like to thank Brian Nyberg, Vera Sit, and Brenda Taylor for involving us in this project. Furthermore, we benefitted from review comments from Susan Chen, Darrell Errico, George Harper, Peter Ott, Paul Reschler, William G. Roland, Vera Sit, Dr. Tim Swartz, and Brenda Taylor. Figure 7.2 is borrowed from a diagram used by Judith Anderson in a talk about probability. The figures were prepared by Courtney Walls.

References

Berger, J.O. 1985. Statistical decision theory and Bayesian analysis. 2nd ed. Springer-Verlag, New York, N.Y.

Berger, J.O. and T. Selke. 1987. Testing a point hypothesis: the irreconcilability of P-values and evidence. J. Am. Statist. Assoc. 82:112–39.

Berry, D.A. 1996. Statistics: A Bayesian perspective. Duxbury Press, Belmont, Calif.

Box, G.E.P. and G.C. Tiao. 1973. Bayesian inference in statistical analysis. J. Wiley, New York, N.Y.

Cox, D.R. and D.V. Hinkley. 1982. Theoretical statistics. Chapman and Hall, New York, N.Y. Reprint.

Dennis, B. 1996. Discussion: Should ecologists become Bayesians? Ecol. Applic. 6:1095–103.

Dixon, P. and A.M. Ellison. 1996. Bayesian inference - introduction: ecological applications of Bayesian inference. Ecol. Applic. 6:1034–5.

Edwards, D. 1996. Comment: The first data analysis should be journalistic. Ecol. Applic. 6:1090–4.

Ellison, A.M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecol. Applic. 6:1036–46.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. Bayesian data analysis. Chapman and Hall, New York, N.Y.

Kass, R.E. and A.E. Raftery. 1995. Bayes factors. J. Am. Statist. Assoc. 90:773–95.

Kennedy, P. 1985. A guide to econometrics. 2nd ed. MIT Press, Cambridge, Mass.

Ludwig, D. 1996. Uncertainty and the assessment of extinction probabilities. Ecol. Applic. 6:1067–76.

Peterman, R.M. and C. Peters. [n.d.]. Decision analysis: taking uncertainties into account in forest resource management. This volume.

Raftery, A.E. 1994. Bayesian model selection in social research (with discussion). *In* Sociological Methodology. P.V. Marsden (editor). Blackwell, Cambridge, Mass. pp. 111–95. first ed.

Swindel, B.F. 1972. The Bayesian controversy. U.S. Dep. Agric. For. Serv. Res. Pap. SE-95. Southeastern For. Exper. Sta., Asheville, N.C.

Wonnacott, T.H. and R.J. Wonnacott. 1977. Introductory statistics. 3rd ed. J. Wiley, New York, N.Y.

References for Some Example Applications

Burk, T.E. and A.R. Ek. 1987. Prior information reduces the error in timber sale appraisals. N. J. Appl. For. 4:3.

Ek, A.R. and J.N. Issos. 1978. Bayesian theory and multi-resource inventory. *In* Proc. Integrated inventories of renewable natural resources workshop, Biometrics, U.S. Dep. Agric. For. Serv., Gen. Tech. Rep. RM-55, pp. 291–8.

Gertner, G. 1987. A procedure to evaluate sampling schemes for improving already calibrated models. For. Sci. 33:632–43.

Green, E.J., M. Kohl, and W.E. Strawderman. 1992. Constrained empirical Bayes estimates for cell values in a two-way table. Can. J. For. Res. 22:1983–7.

Green, E.J. and W.E. Strawderman. 1992. A comparison of hierarchical Bayes and empirical Bayes methods with a forestry application. For. Sci. 38:350–66.

Johnson, D.H. 1989. An empirical Bayes approach to analyzing recurring animal surveys. Ecology 70:945–52.

Taylor, B.L., P.R. Wade, R.A. Stehn, and J.F. Cochrane. 1996. A Bayesian approach to classification criteria for spectacled eiders. Ecol. Applic. 1077–89.

Ver Hoef, J.M. 1996. Parametric empirical Bayes methods for ecological applications. Ecol. Applic. 1047–55.

Wolfson, L.J., J.B. Kadane, and M.J. Small. 1996. Bayesian environmental policy decisions: two case studies. Ecol. Applic. 1056–66.

Appendix 1: Demonstration of Bayes' Theorem using the silviculture example

This appendix uses the example described in Section 7.3 to demonstrate the validity of Bayes' Theorem. The relevant numbers for that example are summarized in Table A1.1

The data in Table A1.1 were previously used to calculate interesting probabilities such as the prior probability that the cutblock is *NSR*:

$$\mathrm{p}(\theta{=}NSR) = \pi_o = \frac{840}{1000} = 0.84\ ;$$

the probability of observing an understocked plot when the cutblock is *NSR*:

$$p(X{=}US|\theta{=}NSR) = \pi_1 = \frac{672}{840} = 0.80\ ;$$

and the probability of observing an understocked plot when the cutblock is *SR*:

$$p(X{=}US|\theta{=}SR) = \pi_2 = \frac{72}{160} = 0.45.$$

Note that there are other interesting probabilities to calculate from Table A1.1 For instance, the probability that both $X = US$ and $\theta = NSR$ occur is known as the joint probability and is denoted by

$$p(\theta{=}NSR,\ X{=}US) = \frac{672}{1000} = 0.672.$$

A general rule is that the joint probability is the product of the prior probability (a marginal probability because it is calculated from the margins of the table) and the conditional probability of the data, *X*, given the "true" state of nature, denoted, mathematically by:

$$p(\theta{=}NSR,\ X{=}US) = p(\theta{=}NSR) \times p(X{=}US|\theta{=}NSR). \tag{A1.1}$$

From Table A1.1, $p(\theta = NSR) = \pi_o = 0.84$ and $p(X{=}US|\theta{=}NSR) = \pi_1 = 0.80$ so that their product is:

$$p(\theta{=}NSR,\ X{=}US) = 0.840 \times 0.80 = \pi_o \times \pi_1 = 0.672.$$

Because the order does not matter, equation (A1.1) can be rewritten as:

$$p(\theta{=}NSR,\ X{=}US) = p(X{=}US) \times p(\theta{=}NSR\,|X{=}US), \tag{A1.2}$$

where $p(X = US)$ is the probability that the one plot will be found to be *US*, and $\mathrm{p}(\theta{=}NSR|X{=}US)$ is the probability that the cutblock is really *NSR* if the plot is observed to be *US*. Notice that this last probability is known as the posterior probability and is what we want to determine from the sampling. It is the probability for a state of nature given our observed data. Relation (A1.2) can be confirmed from Table A1.1 by noting that $p(X{=}US) = 744/1000$ and that $p(\theta{=}NSR\,|\,X{=}US) = 672/744$ so that:

$$p(\theta{=}NSR,\ X{=}US) = 0.744 \times 0.80 = 0.672.$$

Equations (A1.1) and (A1.2) can be set equal to each other:

$$p(\theta{=}NSR) \times p(X{=}US\,|\,\theta{=}NSR)$$
$$= p(X{=}\ US) \times p(\theta{=}NSR\,|\,X{=}US).$$

This equation can be rearranged to obtain a relationship for the posterior probability:

$$p(\theta{=}NSR\,|\,X{=}US) = \frac{p(\theta = NSR) \times p(X = US|\theta = NSR)}{p(X = US)}. \tag{A1.3}$$

In more general (and more readable) terms this equation can be written as:

$$p(\theta|X) = \frac{p(\theta) \times p(X|\theta)}{p(X)}. \tag{A1.4}$$

TABLE A1.1 *Numbers of previously sampled plots (observed as* US *or* S*) from both* NSR *and* SR *cutblocks. Parameters for the prior probability distribution and the two probability models are also shown.*

Probability parameters		Joint distribution (probability model parameters)	
Cutblock is	Prior probability: *p(θ)*	*X = US*	*X = S*
θ = NSR	840 plots (π_0 = 0.84)	672 plots (π_1 = 0.80)	168 plots (1–π_1 = 0.20)
θ = SR	160 plots (1–π_0 = 0.16)	72 plots (π_2 = 0.45)	88 plots (1–π_2 = 0.55)
Total	1000	744	256

This relationship, known as Bayes' theorem, forms the core of the Bayesian statistics methodology. We can confirm that this relationship is true for the example by using the values from Table A1.1 to calculate:

$$p(\theta{=}NSR|X = US)$$
$$= \frac{840/1000 \times 672/840}{744/1000} = \frac{672}{744} = 0.903.$$

Note that this result agrees with that obtained directly from the first column of data in Table 1 and discussed thereafter.

Equation (A1.3) can be written in words as:

(*The posterior probability of a true state of nature given the data) = (the prior probability of that true state of nature) times (the likelihood of the observed data given that true state of nature) divided by (the probability of observing the data*).

Notice that this definition is a more detailed version of equation (1) in Section 7.2. The relationship between the components of Bayesian statistics was presented pictorially in Figure 7.1.

In general, the denominator in equation (A1.3), the marginal probability $p(X{=}\ US)$, can be calculated by summing all the possible values for the numerator of equation (A1.3). For the example, this calculation is:

$$p(X = US) = p(\theta = NSR)\ p(X{=}US|\theta{=}NSR)$$
$$+\ p(\theta{=}SR)\ p(X{=}US|\theta{=}SR)$$

or

$$p(X = US) = \pi_o \times \pi_1 + (1{-}\pi_o) \times \pi_2,$$

so that numerically,

$$p(X = US) = 0.84 \times 0.80 + 0.16 \times 0.45 = 0.744.$$

Thus, for the example, equation (A1.3) can be written as:

$$\text{posterior probability} = p(\theta = NSR|X = US)$$
$$= \frac{\pi_o \times \pi_1}{\pi_o \times \pi_1 + (1{-}\pi_o) \times \pi_2.}$$

Appendix 2: Calculations using several sampling plots for the cutblock

In this appendix we will calculate the posterior probability that the cutblock is *NSR* (not satisfactorily restocked) given that it has been sampled with several plots. While the probabilities remain the same ($\pi_o = 0.84$, $\pi_1 = 0.80$, and $\pi_2 = 0.45$), the probability models for the data [$p(X = US|\theta = NSR)$ and $p(X = US|\theta = SR)$] are now more complicated.

The number of plots observed to be *US* will be denoted by *X*, with n representing the number of plots sampled. The probability of observing *X* out of *n* plots given a specific state of nature, θ, is given by the binomial[9] distribution:

$$p(X|\theta = NSR) = \binom{n}{X} \pi_1^{\,x} (1{-}\pi_1)^{\,(n-X)}, \text{ and}$$

$$p(X|\theta = SR) = \binom{n}{X} \pi_2^{\,x} (1{-}\pi_2)^{\,(n-X)}. \qquad \text{(A2.1)}$$

Suppose that 12 plots were placed in the cutblock and that 7 of them were found to be *US*. Then the conditional probability for the observed data are:

$$p(X = 7|\theta = NSR) = \binom{12}{7}\ 0.80^7\ (1 - 0.80)^5 = \ 0.053$$

and

$$p(X = 7|\theta = SR) = \binom{12}{7}\ 0.45^7\ (1 - 0.45)^5 = \ 0.149$$

We can use equation A1.4 (in Appendix 1) to calculate the posterior probabilities. The denominator now becomes

$$p(X) = \pi_o \times p(X{=}7|\theta{=}NSR) + (1{-}\pi_o) \times p(X{=}7|\theta{=}SR),$$

which is calculated by:

$$p(X) = 0.84 \times 0.053 + 0.16 \times 0.149 = 0.0685.$$

Thus the posterior probability that the cutblock is *NSR* ($\theta = NSR$) given that 7 of the 12 plots were *US* is:

$$p(\theta = NSR|X = 7) = \frac{\pi_o \times p(X = 7|\theta)}{p(X)} \qquad \text{(A2.2)}$$
$$= \frac{0.84 \times 0.053}{0.0685} = 0.652$$

The posterior probability that the cutblock is *SR* is:

$$p(\theta = SR|X = 7) = 1 - 0.652 = 0.348.$$

The posterior probabilities, $p(\theta = NSR|X)$ for all possible values of *X*, are shown in Table 7.3.

[9] This distribution is described in most standard introductory statistical textbooks. $\binom{n}{X}$ is known as the binomial coefficient and $\binom{n}{X} = \frac{n!}{X!(n-X)!}$. If $X = 7$ and $n = 12$ then this is equal to 792. When $n = 1$, the binomial distribution becomes the Bernouilli.

8 DECISION ANALYSIS: TAKING UNCERTAINTIES INTO ACCOUNT IN FOREST RESOURCE MANAGEMENT

RANDALL M. PETERMAN AND CALVIN N. PETERS

Abstract

For forest resource managers, uncertainties are unavoidable because of natural ecological variability and our imperfect knowledge of ecosystems. Nevertheless, management decisions must be made and actions must be taken. Decision analysis, a quantitative method of evaluating management options, can greatly assist that decision-making process because it explicitly uses information on uncertainties. Although widely used in business, decision analysis is particularly useful for forest management because it accounts for uncertainty about states of nature (e.g., current timber volume per hectare, the slope of the relationship between survival rate of a rare bird species and size of patches of mature stands of trees). Decision analysis, in conjunction with Bayesian statistical methods, permits calculation of the potential outcomes of management actions, considering each hypothesized state of nature weighted by its probability of occurrence. Given a clear objective, managers can then rank their management options. A sensitivity analysis can determine how sensitive this ranked order of management options is to different assumptions or parameter values. Sensitivity analysis can also identify research priorities and help resolve conflicts between interest groups about objectives or beliefs about how a forest ecosystem works. Decision analysis is particularly appropriate for the planning stage of an active adaptive management initiative because it can compare the expected performance of different proposed experimental plans, taking into account various uncertainties. This procedure can help identify the best experimental design for an adaptive management plan, as well as its associated monitoring program.

8.1 Introduction

As noted in Nyberg (this volume, Chap. 1) uncertainties are pervasive in natural resource management. Our knowledge of ecosystems is incomplete and imperfect, which creates imprecision and bias in data used to quantitatively describe the dynamics of these systems. Despite the presence of these uncertainties, decisions must be made and regulations must be developed. One purpose of this chapter is to discuss why it is important for decision-makers to explicitly consider uncertainties when evaluating possible management actions, including different designs of adaptive management plans or monitoring programs. Another purpose is to describe decision analysis, a formal, quantitative method that helps decision-makers take uncertainties into account in analyses of options by breaking down the decision problem into tractable components. Several examples will illustrate the benefits and limitations of decision analysis.

8.2 Sources of Uncertainty

Several sources of uncertainties exist in management of forest ecosystems. First, natural variability over space and time is inherent in ecological processes. For example, growth rates of trees and animals may differ among sites, years, and individuals. Such natural variability makes it difficult to forecast responses of ecological systems to different management actions with accuracy or precision. Variability in human behaviour also makes it difficult to forecast how human harvesters and industry will respond to management regulations. Second, further uncertainty exists in data because sampling techniques imperfectly estimate quantities such as density of a certain bird species in a forest, volume of merchantable timber present per hectare, or natural mortality and reproductive rates of mammals. These methods thus create further imprecision and bias in estimates of quantities that vary naturally. Therefore, managers will forecast imperfectly, making it more difficult to achieve a given management objective. Third, management objectives are frequently uncertain, either because they are not well defined or because they change over time. These uncertainties create complications for managers who try to choose the best management option. The challenge for resource managers is how to fully account for the type, direction, and magnitude of uncertainties when making management decisions. One purpose of this chapter is to address this challenge.

Forest managers must recognize that they are not alone in dealing with uncertain systems; uncertainties are present in all natural systems, not just biological ones. For example, we now take for granted the values of several fundamental physical constants such as the speed of light and the charge of an electron, but their estimated values have changed dramatically over time as new experimental techniques emerged (Figure 8.1). It is unsettling to note that several estimates of these physical constants even changed to values well outside the confidence intervals of the previous estimate! Uncertainties due to such measurement biases and errors are likely to be even more pronounced in ecological systems that are relatively variable and complex. Thus, scientists and managers should expect to estimate with error quantities such as the volume of merchantable timber per hectare, abundance of a particular species of cavity-nesting bird, proportion of seedlings surviving per hectare per year, or offspring produced per female mule deer per year. Even if scientists and managers recognize and admit that uncertainties exist, they should not be overconfident about the accuracy or precision of estimated quantities. Because of uncertainties, they cannot expect to know the "right" answer, but should be prepared to use the best estimates along with explicit measures of their uncertainty.

Ecological uncertainties create the potential for making incorrect decisions because they prevent managers from exactly predicting the outcome of a particular management action. When incorrect decisions are made, losses result. In decision theory, a loss is defined as an undesirable outcome of a decision. Losses can be direct losses, such as the elimination of some rare or important species of bird or mammal. Incorrect decisions can also result in opportunity losses when the outcome of the decision is worse than what could have been obtained if the correct decision had been made. For example, an opportunity loss is incurred when a particular thinning regime results in lower net timber revenues than those that could have been generated if a different thinning regime had been implemented. The probability of incurring such losses depends on the degree and type of uncertainty arising from the sources discussed above. Decision theorists define the term "risk" as "expected loss," which is the weighted average loss. This quantity is calculated by multiplying each possible magnitude of loss by a weighting term, which is its probability of occurrence (Morgan and Henrion 1990). To minimize such risks for users as well as management agencies, both scientists and decision-makers should systematically and comprehensively take uncertainties into account. However, this approach is not often taken, as we discuss next.

8.3 Approaches to Making Decisions in the Presence of Uncertainty

Management agencies have historically used several, often ineffective, approaches to making decisions in the presence of uncertainties.

8.3.1 Best estimate approach

One common approach to managing wildlife, forests, and fisheries is to ignore the uncertainties and base management decisions only on the best estimates of all parameters and other quantities. For example, into the 1970s, allowable annual cut (AAC) in British Columbia was calculated with a formula using only the best estimates of parameters, without taking uncertainties into account (Pearse 1976). The problem with this approach is that incorrect estimates can lead managers to make incorrect or suboptimal decisions. Nevertheless, this focus on using the best point estimates is very common, especially where admitting uncertainty would provide user groups with leverage to argue for increased harvests or decreased protection of non-timber values such as wildlife. To avoid such debates, managers sometimes request that scientists only provide them with their best point estimates, even though everyone is aware that uncertainties exist.

8.3.2 Qualitative approach

A second approach to making decisions takes uncertainties into account, but only qualitatively or crudely, rather than rigorously. This approach is manifested in four ways:

1. First are cases where managers use ecological uncertainties to justify maintaining the status quo. For instance, in 1991 the Forest Resources Commission recommended that "the Allowable Annual Cut of Timber Supply Areas or Tree Farm Licenses not be raised or lowered until and unless new timber inventory data and subsequent yield analysis clearly justify an adjustment, except in those obvious cases where current information strongly support a change" (Peel 1991, p. 84, recommendation #87). In other words, the default is to maintain the status quo until uncertainties are clarified to the point where a change in AAC is clearly indicated.
2. Some people have used a qualitative approach to justify extreme pessimism about the response to a management action. For example, the public

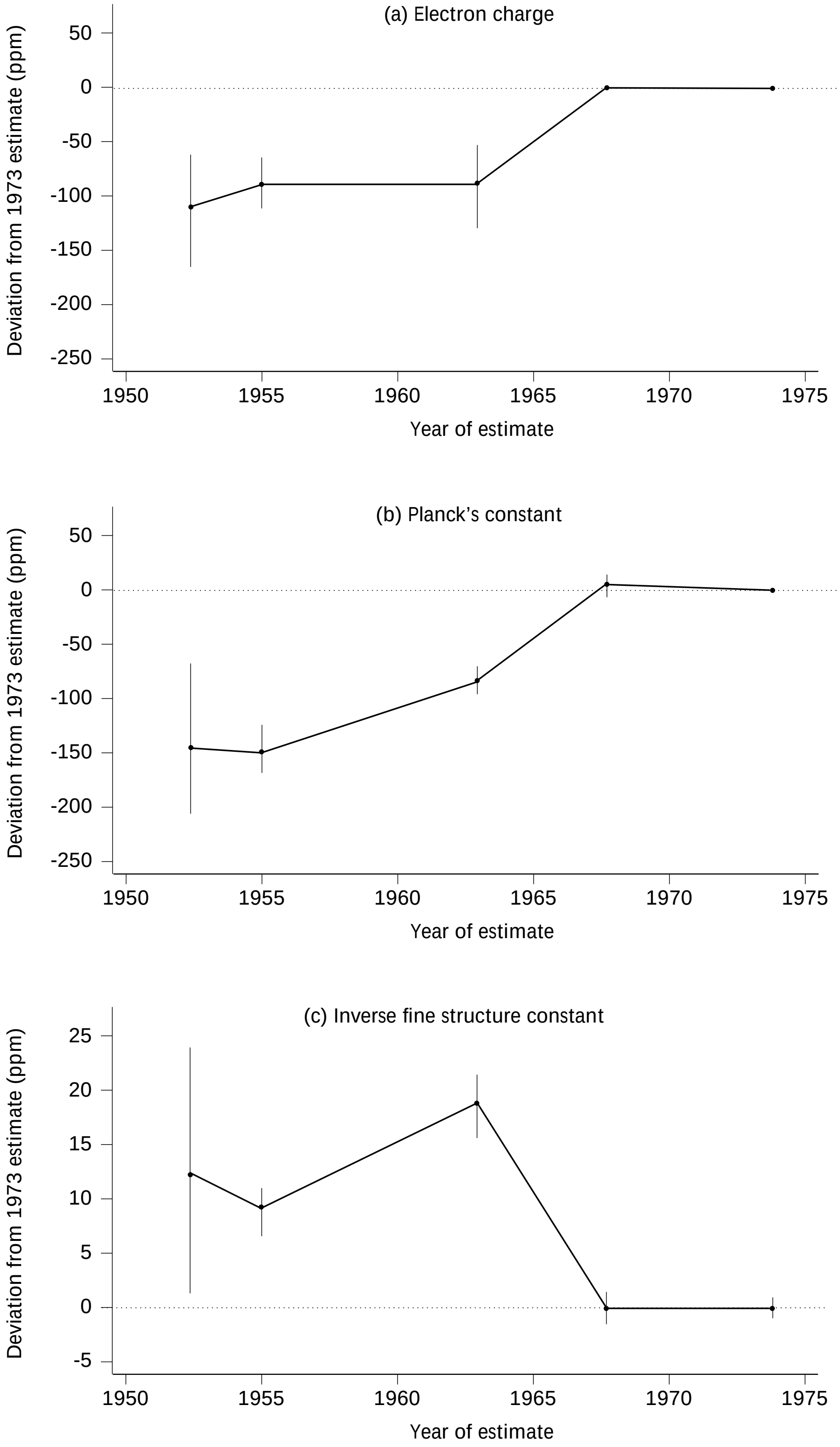


FIGURE 8.1 *Changes in estimates of various physical constants as new experimental or measurement methods were developed. Data points are mean estimates of physical constants and vertical bars represent standard errors of the mean estimates. All values are in units of deviations from their 1973 estimates, in parts per million (ppm). (Adapted from Henrion and Fischhoff 1986.)*

opposition to plans for spraying for gypsy moth in New Westminster, B.C., arose partly because the effects of spraying on public health were uncertain.

3. Uncertainties can also be used qualitatively to justify a moderately pessimistic outlook and to implement a conservative approach to management. For example, when choosing the density of lodgepole pine seedlings to replant, the density is often increased by some arbitrary amount to allow for uncertainty in tree mortality due to attack by pests (Errico 1989).

4. Finally, resource users or managers may use uncertainties qualitatively to justify an optimistic view of how systems will respond to management. Many cases in forestry and in fisheries indicate that industry has used uncertainties in this way to promote increased harvests or reduced protection of the resource. For example, harvest rates of forests have been increased in some areas in part based on optimistic predictions about the future effects of thinning and other enhanced silvicultural activities on timber supply.

These four qualitative approaches to considering uncertainty in decision-making may result in either unnecessarily restrictive or excessively lenient policies because the effects of uncertainties on the outcomes of management actions are not considered quantitatively and explicitly.

8.3.3 Quantitative approach

A third and better approach to making decisions is to take uncertainties into account quantitatively by considering a range of possible responses of an ecological system to each management action. In doing so, managers can select the option that minimizes the risk. For example, several possible growth responses of trees to specific amounts of thinning could be explicitly considered to reflect uncertainty in that response, rather than choosing just one of those responses as the sole basis for the decision.

Several methods are available for taking uncertainties into account quantitatively, including decision analysis, Monte Carlo simulation, and formal optimization techniques. Decision analysis was developed in the 1960s in business (Raiffa 1968) to help decision-making in the presence of economic uncertainties, which is directly analogous to making decisions in the presence of ecological uncertainties. Decision analysis is becoming a popular tool in resource management (e.g., Lord 1976; Walters 1981, 1986; Cohan et al. 1984; Parkhurst 1984; Bergh and Butterworth 1987; Holbert and Johnson 1989; Parma and Deriso 1990; McAllister and Peterman 1992a; McDaniels 1992; Thompson 1992; Hilborn et al. 1994; Maguire and Boiney 1994; Reckhow 1994; Adkison and Peterman 1996). This popularity is due to several reasons.

First, most problems in resource management are too complex (with lags, nonlinearities, threshold phenomena, and cumulative effects) to permit the use of formal optimization techniques (see Clark 1990 for some exceptions). Second, decision analysis can help managers rank proposed management actions based on quantitative assessments of probabilities of uncertain events and the desirability of possible outcomes (Keeney 1982; Howard 1988; Clemen 1996). Decision analysis can be thought of as one type of risk assessment in that it considers the uncertainties that create risks. Although decision analysis cannot guarantee that a correct decision will be made each time, it will improve the quality of several similar decisions over the long term because it explicitly takes uncertainties into account quantitatively (Von Winterfeldt and Edwards 1986). Similarly, taking the optimal action identified by a decision analysis does not guarantee a certain desired outcome, but it increases the probability of a desirable outcome occurring. Finally, decision analysis can combine Bayesian statistical analysis and stochastic models (Monte Carlo simulations) into a structured, systematic approach to making decisions. Complex decision problems are broken down into smaller and more manageable components; these components are then recombined to determine the optimal action. This process makes decision analysis a useful tool for decisions involving complex ecological and human responses to management actions, which certainly characterize forest management.

8.4 Eight Components of Decision Analysis

To make a complex decision problem in forestry more tractable, decision analysis breaks the problem down into eight components:

1. management objectives;
2. management options;
3. uncertain states of nature;
4. probabilities on the uncertain states of nature;

TABLE 8.1 *A generalized decision table showing calculation of expected outcomes for two potential management actions, given two possible states of nature (Hypothesis 1 and 2) with their associated probabilities (P_1 and P_2). Compare with Figure 8.2.*

Hypotheses or uncertain states of nature	Probabilities	Potential management action #1	Potential management action #2
Hypothesis 1	Probability that Hypothesis 1 is correct (P_1)	Consequence of action 1 if Hypothesis 1 is correct (C_{11})	Consequence of action 2 if Hypothesis 1 is correct (C_{21})
Hypothesis 2	Probability that Hypothesis 2 is correct (P_2)	Consequence of action 1 if Hypothesis 2 is correct (C_{12})	Consequence of action 2 if Hypothesis 2 is correct (C_{22})
		Expected consequence of action 1 = $(P_1 \times C_{11})+(P_2 \times C_{12})$	Expected consequence of action 2 = $(P_1 \times C_{21})+(P_2 \times C_{22})$

5. model to calculate the outcome of each management action for each state of nature;
6. decision tree or decision table;
7. ranking of management actions; and
8. sensitivity analyses.

A generalized decision table (e.g., Table 8.1) can be used to structure the decision analysis of simple problems. In this table, two alternative management actions are listed across columns and alternative hypotheses or uncertain states of nature, with their associated probabilities (P_1 and P_2), are placed in rows. For each combination of action and hypothesis, the consequences or outcomes (C_{11}, C_{12}, etc.) are calculated using a model. The "expected" value of the consequence for a particular management action (last row) is then calculated from the weighted average of all possible consequences for that action, where the weighting is the probability of the hypothesis that gives rise to each consequence.

For more complex problems, a decision tree can be used to structure the analysis (Render and Stair 1988; Clemen 1996). The generalized decision tree in Figure 8.2 corresponds to the decision table in Table 8.1. Alternative management actions in Figure 8.2 are represented by branches emerging from a square

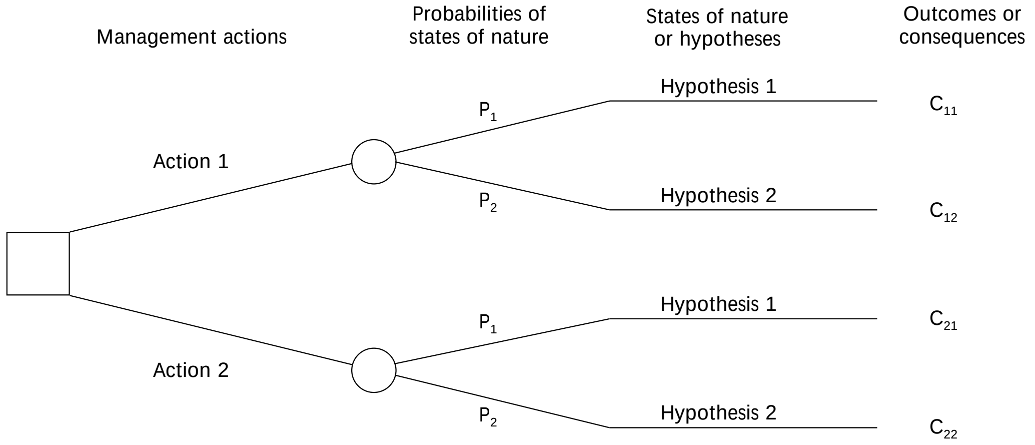


FIGURE 8.2 *A simple example of a generalized decision tree showing two different management actions and two possible states of nature (Hypothesis 1 and 2) with their associated probabilities (P_1 and P_2). The square at the left is the "decision node" and the circles are "chance nodes." The consequences associated with each combination of management action, i, and state of nature, j, are designated C_{ij}. This decision tree is the graphical equivalent of the general decision table shown in Table 8.1.*

decision node, and uncertain states of nature or hypotheses are represented as branches coming from the circular chance nodes. The probability of each uncertain state of nature is shown explicitly for each state-of-nature branch. Outcomes or consequences of each management action, given each state of nature, are shown on the right. Decision trees can accommodate much more complexity than a decision table by including numerous branches and uncertainty nodes.

We will use an application of decision analysis to forest management in Tahoe National Forest, California (Cohan et al. 1984) to illustrate the eight components of this method. The purpose of Cohan et al.'s particular decision analysis (referred to as the "Tahoe example") was to determine what treatment should be applied before a prescribed burn on a recently harvested forest site. Figure 8.3 shows the decision tree for this problem; its components are explained below.

8.4.1 Management objectives

Decision analysis requires a clearly defined management objective or goal so that the different management actions can be ranked by how well they are expected to attain the objective. The objective is usually stated explicitly in terms of maximizing (or minimizing) one or more quantitative measures of performance (such as expected value of future timber harvests). However, decision analysis can also accommodate situations in which the objective is to choose an action that produces a performance measure, such as abundance of some rare bird species, that is within an acceptable range of values. In this case, actions that do not lead to outcomes within this range can be discarded, and some secondary criterion (such as minimizing cost) can be used to choose from the remaining actions. As emphasized by Keeney (1992), identifying objectives requires carefully applying various procedures to ensure, for instance, that "fundamental" objectives are not confused with the means needed to attain them.

In the Tahoe example (Figure 8.3), the management objective was to maximize the expected net resource value of the forest following the prescribed burn. That value took into account the value of the timber harvested, as well as the cost of carrying out the pre-burn treatment (if any), the cost of the prescribed broadcast burn, and the cost incurred from an escaped fire (if one escaped).

In the case of British Columbia's forests, management objectives can involve timber value, recreational use, wildlife habitat, and quality of "viewscapes" in various combinations and with various relative importances. For example, a primary management objective in Clayoquot Sound is to maintain long-term productivity and natural diversity of the area. Subgoals include maintaining watershed integrity, biological diversity, and cultural, scenic, recreational, and tourism values (Scientific Panel for Sustainable Forest Practices in Clayoquot Sound 1995).

8.4.2 Management options

Managers need to define a list of alternative actions from which to choose the best option. Considerable thought should be put into developing innovative options, as well as into identifying feasible ones (Keeney 1982).

The Tahoe prescribed burn problem has two alternative management actions. These alternatives are shown in Figure 8.3 as two branches emerging from the square "decision node." One choice was to conduct the prescribed broadcast burn without any pre-burn treatment of the site ("burn only"). The other alternative was to pile up timber slash from the clearcut harvest before the broadcast burn ("YUM and burn"). This latter treatment, referred to as yarding unmerchantable material (YUM), incurs additional costs but reduces the probability of fire escaping and increases the chances of a successful burn. Cohan et al.'s (1984) question was, "Is YUM worth the additional cost?"

8.4.3 Uncertain states of nature

Uncertain states of nature are parameters or quantitative hypotheses that are treated explicitly as uncertainties in an analysis, usually by considering a range of values for one or more parameters in a model (see Section 8.4.5). Such uncertain parameters lead to a corresponding range of forecasts of outcomes of management actions. For instance, it may be difficult to estimate the effect of different sizes of "leave patches" in a retention harvesting strategy on abundance of a bird population because of uncertainty about how the probability of blowdown is affected by patch size (i.e., whether that probability is a steeply rising function of patch size or a relatively flat one). There is also considerable uncertainty about the benefits of some requirements in the British Columbia Forest Practices Code for meeting objectives related to biodiversity or recreational use. For example, it is unclear whether the survival rate of juvenile coho salmon is greatly or only slightly affected by the width of riparian forest that the Code requires to be left along stream banks.

Two major uncertainties in the Tahoe example (Figure 8.3) involved the fire behaviour and the magnitude of costs associated with what Cohan et al. (1984) referred to generally as "problem" fires. Uncertainty in fire behaviour was represented by defining three types of fires: a successful burn, a "problem" burn, and an escaped fire. The second uncertainty was the cost of a "problem" burn (high, intermediate, or low cost). These uncertain states of nature are shown as branches emerging from circular "chance nodes" in Figure 8.3.

8.4.4 Probabilities on the uncertain states of nature

In forest management, considerable uncertainty usually exists about states of nature such as those listed in Section 8.4.3 because of short data series, natural variability, and measurement error and bias. However, scientists and decision-makers need to state a relative "degree of belief," or probability, for these different states of nature so that they can forecast the expected outcome of each possible management action and determine its ranking. For example, wide confidence limits on the slope of a relationship between survival rate of trees to a given age and initial stocking (density of seedlings) can produce a range of forecasts about future harvests from stands that are

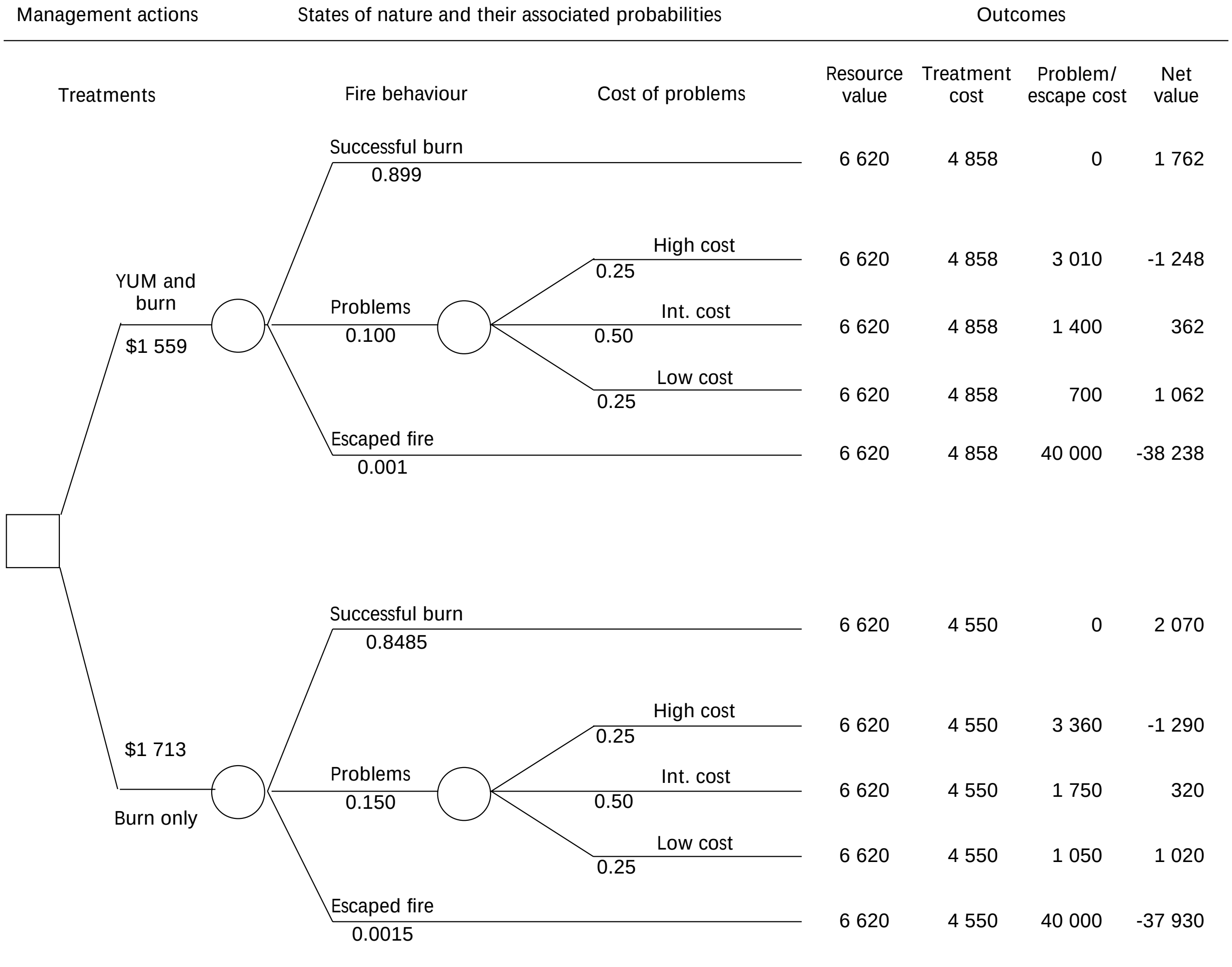


FIGURE 8.3 *Decision tree for the example described in the text for the Tahoe National Forest. The management options (treatments) are to "burn only" or "YUM and burn"; the latter refers to "yarding unmerchantable material," where the slash material from the logging operation is piled up before burning. Outcomes are costs in dollars for a 14-acre site. The resulting expected net resource values for each management option are indicated next to the option. See text for details. (Adapted from Cohan et al. 1984.)*

replanted at a specific density. In this case, managers need a probability for various slopes of that relationship to estimate the expected harvest levels for different stocking densities.

Unfortunately, classical statistics do not provide such probabilities. Most ecologists describe uncertainty in estimates of states of nature (e.g., slopes or other parameters) with standard errors, confidence limits, and coefficients of variation. They also routinely apply classical statistical inference methods to test point null hypotheses. However, such procedures are inadequate for decision-making for the following reasons.

First, hypothesis tests are too restrictive for making decisions in the presence of uncertainties because they only provide information relevant to two states of nature: the null hypothesis and a specified alternative. In hypothesis testing, a point null hypothesis, H_O, (e.g., the slope of the relationship between average volume per tree at a given age and initial density = 0) is tested with data by some classical method such as a *t*-test. The null hypothesis is either rejected in favour of the alternative hypothesis, H_A, or it is not, based on a comparison of the computed P-value with the pre-determined α. For two reasons, this dichotomous approach to describing the state of nature (H_O versus H_A) is inappropriate to describe ecological uncertainty in decision analyses. First, managers need to consider several different H_A estimates of the slope as possible states of nature, not just Ho and a single H_A, because different slopes may have very different implications for the selection of an initial density of seedlings to replant (e.g., Figure 8.4). Second, the P-value resulting from a standard hypothesis test refers to the probability of obtaining the test-statistic by chance alone if the H_O were true and does *not* state the probability that H_O or any other possible hypothesis is correct. Therefore, P-values do not provide the decision analyst with the required probability for even one state of nature (Berger and Berry 1988), let alone several. Thus, the classical statistical approach to hypothesis testing is not a useful framework for characterizing ecological uncertainties as input to decision analyses as described here.

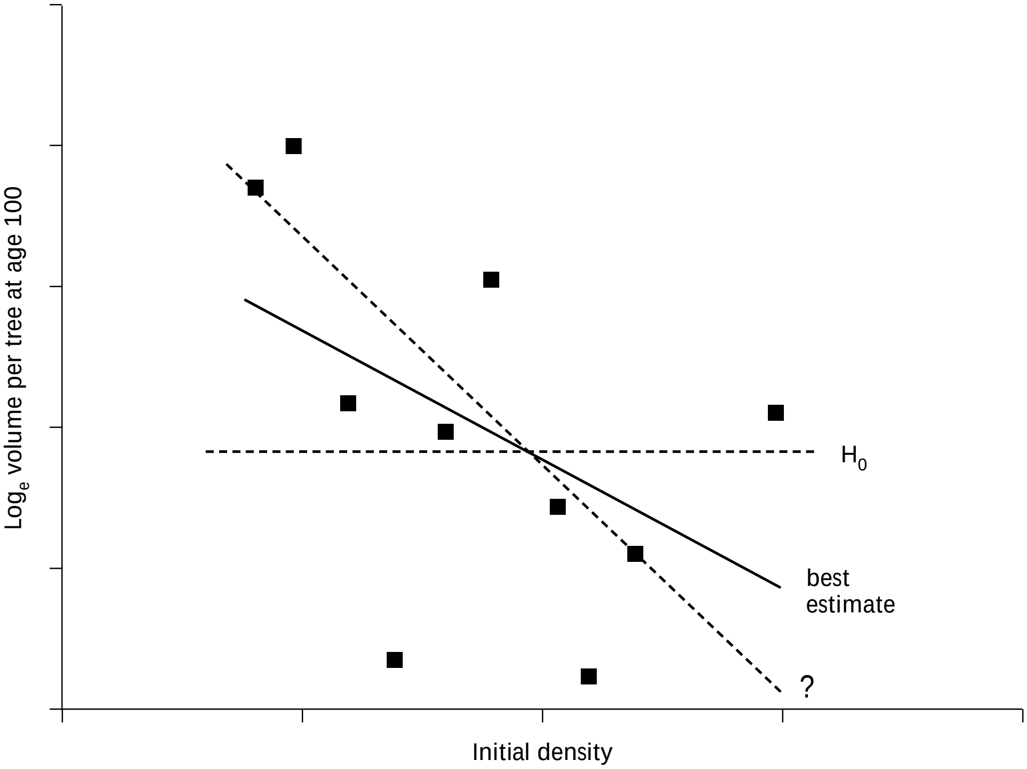


FIGURE 8.4 *Possible models for hypothetical data on average volume per tree at age 100 years as a function of initial density. The solid line is the best-fit regression line; dashed lines represent other possible, but less likely, hypotheses about the true underlying relationship, including the null hypothesis, H_O, of no relationship.*

For similar reasons, standard errors of parameter estimates and 95% confidence limits are also not useful characterizations of uncertainties for making decisions. Specifically, they do *not* indicate the probability to be placed on *different* possible states of nature, even though scientists commonly make this misinterpretation about confidence intervals (e.g., see Sokal and Rohlf 1969, p. 142; Bergerud and Reed, this volume, Chap. 7). Some readers might consider using statistical power analysis for this purpose, which is the probability of correctly rejecting H_O (using a particular method such as a *t*-test) when a specific alternative hypothesis is true. While power analysis is indeed useful for designing experiments and monitoring programs that will reduce uncertainty (e.g., Peterman 1990a, 1990b; Osenberg et al. 1994; Mapstone 1995), statistical power does *not* indicate the probability that a given H_A might be true in nature. Thus, statistical power analysis does not provide sufficient information to decision-makers about the relative degree of belief in alternative states of nature.

Instead, quantifying uncertainties in states of nature requires an assessment of "*the relative merits of rival hypotheses* in the light of observational or experimental data that bear upon them..." (Edwards 1992, p. 1; emphasis ours). Three techniques are available to do this. First, long and detailed historical data sets can provide information about the relative frequency of events. For example, historical data on forest fires can provide probabilities for different intensities and sizes of forest fires in a specific region. Similarly, data from stream gauges can quantify the probability of different heights of streams. Unfortunately, such lengthy continuous records are not common. Second, where data are inadequate, estimates of probabilities for different states of nature can be elicited from experts using various techniques, based on their knowledge of the system under study (Morgan and Henrion 1990). Third, Bayesian statistical analysis is appropriate where some but not enough data are available to generate frequency distributions as discussed in the first technique. Bayesian statistical methods use these data to generate probabilities representing degrees of belief for different values of parameters (see Bergerud and Reed, this volume, Chap. 7). This approach uses Bayes' theorem to calculate the posterior probability that a particular hypothesis is correct, given the data and some prior probability distribution of the hypotheses based either on other, independent data (when available) or on expert opinion (Box and Tiao 1973; Press 1989). For example, rather than merely considering the two possibilities that the slope of the relationship between volume per tree and initial density in Figure 8.4 is either significantly different from zero or not, a Bayesian statistical analysis would use the observed data to calculate probabilities for a wide range of different slopes (Figure 8.5), each of which has different implications for the management decision of how many seedlings to replant. Thus, the output of Bayesian analyses provides exactly the information required for a decision analysis: the probabilities associated with each of several uncertain states of nature. Ellison (1996) presents an easy-to-read introduction to use of Bayesian statistics in applied ecology; the Crome et al. (1996) paper in the same issue applies Bayesian statistics to the problem of detecting effects of logging on forest-dwelling birds.

In the Tahoe example in Figure 8.3, forestry staff provided estimates of probabilities of the three types of burns and the three levels of costs associated with problem fires. These probabilities appear on the branches corresponding to the appropriate uncertain state of nature (Figure 8.3). The probabilities placed on the three types of burns were different for the two management options ("YUM and burn," or "burn only," because treating the site before burning increased the probability of a successful burn and reduced the probability of either a problem burn or an escaped fire. Managers estimated that the probabilities of high, intermediate, and low costs of a problem burn (if one occurred) would be the same for both management actions.

8.4.5 Model to calculate outcomes

Another key element of decision analysis is the model used to calculate the consequences of each combination of a particular management action and each possible state of nature. The bounds, assumptions, and complexity of the model will depend on the availability of data. Whatever type of model is used, it must produce quantitative indicators of the consequences of each alternative management action, such as revenue from timber or an index of bird diversity. Those indicators must relate directly to the management objective stated in the first component of decision analysis.

In the Tahoe example, the relative costs and timber revenues resulting from different treatments were estimated using models of fire behaviour, fire effects, and economics developed by timber management staff in Tahoe National Forest. The forecast outcomes of these models for each alternative management action and for each uncertain state of nature are shown on the right side of Figure 8.3. The net resource value is the resource value minus the treatment cost and the cost of a problem fire or escaped fire. For example, the simulated net resource value of the "YUM and burn" option, if a "successful burn" resulted, was $1762 on this 14-acre site. For the same action, but assuming that a problem fire recurred that had high costs, their simulated net resource value was -$1248.

8.4.6 Decision tree or decision table

A decision tree or decision table provides a logical framework for ranking the different management actions by combining the states of nature, their probabilities of occurrence, and their outcomes. These rankings are based on "expected values" of outcomes, or weighted average outcomes, for each action. That is, each outcome is weighted by the probability assigned to the associated state of nature (parameter value or hypothesis). Summing these weighted outcomes for each management action gives the expected value of that action. Thus, the expected value as defined by decision theorists represents the weighted average quantity, *not* necessarily the specific value that you would expect to see in the short term (Lindley 1985). The latter is in principle unknowable, given the uncertainty.

The decision tree for the Tahoe problem (Figure 8.3) illustrates this structure. For each management action (type of pre-burn treatment), each possible behaviour of fire, and each possible level of cost of problem fires, there is a resulting value of the timber resource, a treatment cost, and a cost resulting from problem or escaped fires. The expected value of each alternative action is the sum of the net resource value for each state of nature, multiplied by the probability of that state occurring. Thus the "YUM and burn" alternative has an expected value

$$\begin{aligned} EV &= (0.899 \times 1762) + (0.1 \times 0.25 \times (-1248)) \\ &\quad + (0.1 \times 0.5 \times 362) \\ &\quad + (0.1 \times 0.25 \times 1062) \\ &\quad + (0.001 \times (-38\ 238)) \\ &= \$1559. \end{aligned}$$

By similar calculation, the expected value of the option without a pre-burn treatment is $1713.

Although the probability of an escaped fire is very low (0.001 or 0.0015), its cost could contribute significantly to the total *expected* net resource value of each management option. This example shows how even low-probability events may affect which action is optimal, if the costs of such events are large enough.

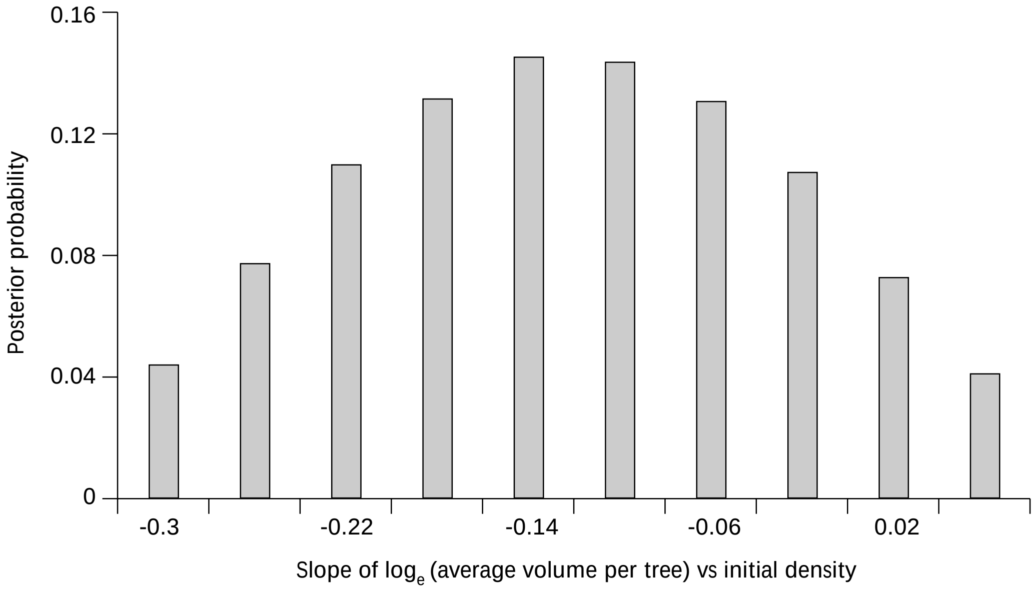


FIGURE 8.5 *Posterior probabilities for different slopes of a linear model for the hypothetical data shown in Figure 8.4. Posterior probabilities were calculated using Bayesian statistics. The best-fit line shown in Figure 8.4 has the highest posterior probability, but other lines with different slopes also have reasonably high probabilities. These probabilities can be used in a decision analysis to represent the relative degree of belief in the different slopes.*

8.4.7 Ranking of management options

The management actions are ranked by applying the management objectives identified in the first component of decision analysis. For instance, if the objective is to maximize the expected value of outcomes, each management action can be ranked using the calculations from the decision tree. In the Tahoe example in Figure 8.3, the optimal action is to "burn only" without treating the site beforehand; its expected value of $1713 was greater than the $1559 for the other option. This "burn only" option maximizes the expected net resource value, even though the probabilities of a problem fire or escaped fire are higher with this alternative than with the "YUM and burn."

By ranking management options in this way, decision analysis explicitly considers uncertainties by taking into account the probability that different states of nature may exist, as well as their effect on the expected outcomes of each management action. The optimal decision identified when uncertainties are used in this manner is referred to as the Bayes decision (Morgan and Henrion 1990).

8.4.8 Sensitivity analyses

Decision analysis provides only one possible answer to a decision problem because the optimal decision may depend on the assumptions made, the value of various parameters in the model, the structural form of relationships in the model, or the probabilities placed on the states of nature. Therefore, managers must also be given results of sensitivity analyses, which directly show how the rank order of management actions (i.e., the best decision) is affected by these assumptions. If such analyses show that a given action is still optimal over a wide range of assumptions, then such assumptions can be deemed relatively unimportant and managers will be confident that the recommended action is indeed the best one. However, if the rank order of management actions is sensitive to different assumptions, then more data must be collected for that particular parameter or assumption. In this manner, a sensitivity analysis of a decision analysis can identify future research priorities.

Although Cohan et al. (1984) did not conduct a formal sensitivity analysis of their Tahoe example shown in Figure 8.3, several parameters could affect the optimal decision, including the additional costs of performing the YUM treatment, the costs associated with an escaped fire, and the probability of an escaped fire if the YUM treatment is *not* used. To demonstrate sensitivity analysis, we calculated the effect of the last parameter on the optimal decision by repeating the decision analysis using several possible values of the probability of the fire escaping for the "burn only" option. The parameter values investigated ranged from 0.001 to 0.009 (Figure 8.6). Results show that the "burn only" option remained the best option (i.e., generated the largest expected dollar value of the resource) as long as the probability of having an escaped fire was less than 0.0055. However, if that probability was actually 0.0055 or greater, then the "YUM and burn" option became the one with the largest expected dollar value of the resource. Thus, over a certain range of this parameter value, the decision that was optimal in the original baseline case ("burn only") remained optimal, but, outside of that range, the optimal decision switched to "YUM and burn." Such results should be presented to experts to determine whether a value greater than or equal to 0.0055 for the probability of an escaped fire without YUM is within the realm of possibility. If this range is not plausible, decision-makers can be confident that uncertainty in this parameter does not affect their decision. However, if a value in this range is plausible, the value of this parameter in nature becomes important for decision-making, and high priority should be placed on obtaining a better estimate of this probability.

Sensitivity analyses can also be used to show how different management objectives may or may not affect the choice of the optimal decision. This is particularly important when objectives include more than just maximizing the expected value of timber harvested. Diverse objectives of various stakeholder groups are currently commonplace in forest management in British Columbia. For example, objectives in the Kamloops Land and Resource Management Plan also include protection of habitat, maintenance of diverse recreational fishing opportunities, and conservation of Pacific salmon (Westland Resource Group 1995). In this type of situation, a quantitative sensitivity analysis using the method of decision analysis can show how similar to one another objectives would have to be lead to the same management option being chosen (e.g., Maguire and Boiney 1994). In some cases, relatively little change in the objectives of one or more interest groups may lead them to recommend the same action, thereby resolving a conflict.

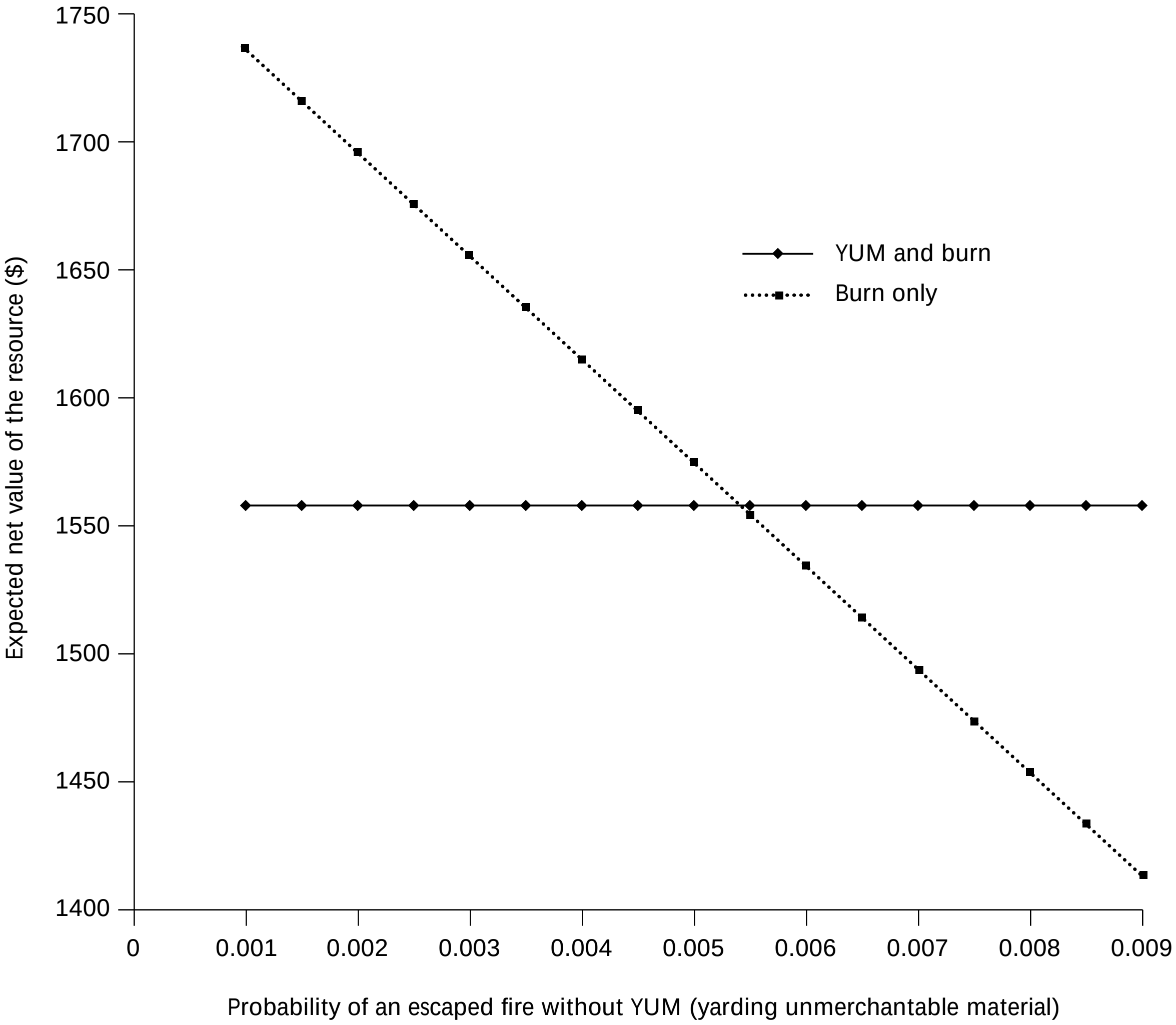


FIGURE 8.6 *An example sensitivity analysis of Cohan et al.'s (1984) decision analysis on the Tahoe burning example (Figure 8.3). Lines show the expected net dollar value of the resource for different probabilities of an escaped fire under the "burn only" option (i.e., without YUM). The solid line represents the "YUM and burn" option; the dashed line is for the "burn only" option. The best estimate provided by forestry staff of the probability of having a fire escape under the "burn only" option was 0.0015 (i.e., 1.5 chances out of 1000), but there is uncertainty in this estimate. The sensitivity analysis shows that the "burn only" option had the highest expected dollar value as long as this probability was less than 0.0055. Above that value, the expected value of the "YUM and burn" option was greater than that of the "burn only" option.*

8.5 Application of Decision Analysis to Adaptive Management

Because management of forests is an uncertain science, Walters (1986) argued that resource managers should manage in an active adaptive manner. In other words, they should carefully design management actions as experiments, just as laboratory experiments or monitoring programs would be designed before their implementation (Hairston [editor] 1989). Well-designed experiments generate rigorous new information about the relative effectiveness of each action or about the different hypotheses about biological processes. Acting adaptively will tend to reduce future uncertainties and thereby improve future management (Peterman and McAllister 1993).

If decision-makers take this approach, they must be able to evaluate alternative management actions, including different adaptive management plans, based on the plans' abilities to generate timely and cost-effective information. The key question is, which experimental design is the most appropriate? Dozens of possible experimental designs could be implemented, not all of which are going to be equally informative, let alone feasible. If a suboptimal design is chosen, the information may be too costly or may not reduce uncertainties. Therefore, decision-makers need some way to compare different experimental designs and identify the one that is expected to maximize the benefits of adaptive management. Decision analysis is an appropriate method to do this.

For instance, suppose that we are planning an experiment such as the ones currently investigating silviculture techniques in British Columbia. Assume that an objective is to maximize timber value. Different randomly selected plots can be thinned to different densities 20 years after replanting to stimulate growth of remaining trees. However, many possible arrangements of treatments exist (Table 8.2 shows only a few examples). The question is, which of these arrangements should be used? Decision analysis can help answer these questions by comparing the expected outcomes of different options, taking uncertainties into account about the amount of release to be experienced by trees in different densities at age 20. In this sense, decision analysis can integrate several of the methods described in previous chapters (e.g., power analysis, sampling, experimental design) and provides a structured way to choose among many possible arrangements of experiments or adaptive management plans.

8.6 Other Examples of Decision Analysis

Many examples are available from fields within resource management where decision analysis or similar methods of accounting for uncertainty have been used to help structure a resource management problem and to identify an optimal action. For instance, in addition to the Tahoe example of pre-burn treatments previously described, Cohan et al. (1984) presented several other cases that applied decision analysis to fire management in U.S. National Forests. In all cases, useful insights into prescribed burning resulted from taking uncertainties into account and breaking the complex problems into understandable components. The authors also noted that using decision analysis to document the rationale for decisions improved the rate of learning by management agencies. Managers involved with silviculture experiments have also used decision analysis to compare the expected performance of different planting, thinning, and fertilization strategies.

Stahl et al. (1994) evaluated different but very important questions for forest managers: what is the optimal method for conducting forest inventories, given that more precise methods cost more, and how often should an inventory be done on a stand? While the researchers used formal optimization methods, their approach was structured much like a decision analysis; they identified uncertainties about the state of nature (current timber volume) when comparing the effects of different inventories on the expected value of net timber revenue (value of timber minus costs of harvesting and conducting inventories). This analysis considered such uncertainties explicitly by assuming that each of three

TABLE 8.2 *Some possible arrangements that could be considered for a thinning experiment. Each arrangement consists of a different number of replicates at various densities of trees, which might be necessary because of logistical constraints.*

	Number of replicate plots at each density		
Density (stems/ha)	Option 1	Option 2	Option 3
250	3	3	4
500	2	3	4
750	3	2	0
Control (unthinned)	2	2	2

inventory methods would produce a probability distribution of estimates of timber volumes at any given time. The inventory methods differed in cost (high, medium, or low) and precision (high, medium, or low).

Stahl et al. (1994) found that, in general, several inexpensive and less precise inventories taken only a few times during the life of a stand resulted in a higher expected net income than a single, expensive but very precise inventory. In addition, the authors concluded that precise inventory information was more valuable when the potential losses in income due to incorrect decisions were large. This conclusion is perhaps intuitive, but Stahl et al. were able to quantitatively estimate the relative value of different methods of doing forest inventories by explicitly considering uncertainties in information.

In wildlife management, Maguire (1986) used decision analysis to recommend an appropriate conservation strategy for Whooping Crane populations to minimize their probability of extinction. Maguire evaluated whether it is better from a biodiversity standpoint to create a single large population or several small ones, given that random catastrophic events can occur (a common debate in conservation biology; see Simberloff and Abele 1976). In the Whooping Crane situation, when Maguire (1986) took the uncertainties associated with severe storms into account, the optimal action was to move some of the Whooping Cranes and create two separate populations. This approach was better than keeping them as a single population that had a higher probability of extinction if a rare severe storm occurred in that one location.

Decision analysis has also been applied to complex land use issues, such as the decision whether to preserve and/or mine in the Tatshenshini area of wilderness in northwestern British Columbia (McDaniels 1992). There, the major uncertainties included the environmental values associated with preserving the area, the tax revenue to be generated by mining, the question of whether mining would actually go ahead given the regulatory process, and other uncertainties. The analysis suggested that preservation of the entire area as a park would have the greatest expected value, taking into account the "nonmarket" value of the wilderness.

Within the field of natural resources, decision analysis has been used most widely in fisheries management. For instance, several authors have used decision analysis to identify optimal management actions for Pacific salmon (e.g., Lord 1976; Walters 1981, 1986). Decision analysis was also able to identify the optimal precautionary safety margin to apply to harvest rates of other marine fish species, given uncertainties in stock abundance and productivity (Frederick and Peterman 1995).

A final fisheries example from the northwestern shelf of Australia (Sainsbury 1988, 1991; Sainsbury et al. 1997) demonstrates particularly well how decision analysis can be used in the design phase of an experimental, or active adaptive management program. Foresters can learn considerably from this case study because it is one of the few large-scale active adaptive management experiments ever implemented, as well as one of the few to use formal decision analysis in the planning stage (also see Walters 1986; McAllister and Peterman 1992b). This case study is therefore worth discussing in detail.

The objectives of this experiment were to determine why the abundances of two economically valuable groups of groundfish species were declining relative to less valuable species and to take appropriate management action (Sainsbury 1988). In 1985, Sainsbury proposed four different hypotheses, or "states of nature," that could potentially explain the historical decrease in abundance of the valuable species relative to the less valuable ones. These hypotheses were an intraspecific mechanism that inhibited the valuable species, two different interspecific interactions between the valuable and less-valuable species that kept the former at low abundances, and a mechanism in which the existing trawl fishery disrupted the preferred ocean floor habitat of the valuable species. Sainsbury proposed five experimental, or active adaptive, management regimes to distinguish among these hypotheses (see W_A to W_E in Figure 8.7, which shows the major elements of Sainsbury's decision analysis). These experimental management strategies ranged from continuing the existing trawl fishery, to stopping the trawl fishery for some period and using a trap fishery only, to several activities in various spatial areas (including no fishing, trap fishing only, trawl fishing only, or both). Sainsbury's decision analysis forecasted the expected economic value of the fish catch for each of these management strategies for each of the four possible "states of nature." These states of nature were weighted by their probability of occurrence (P_1 to P_4), as derived from historical data and

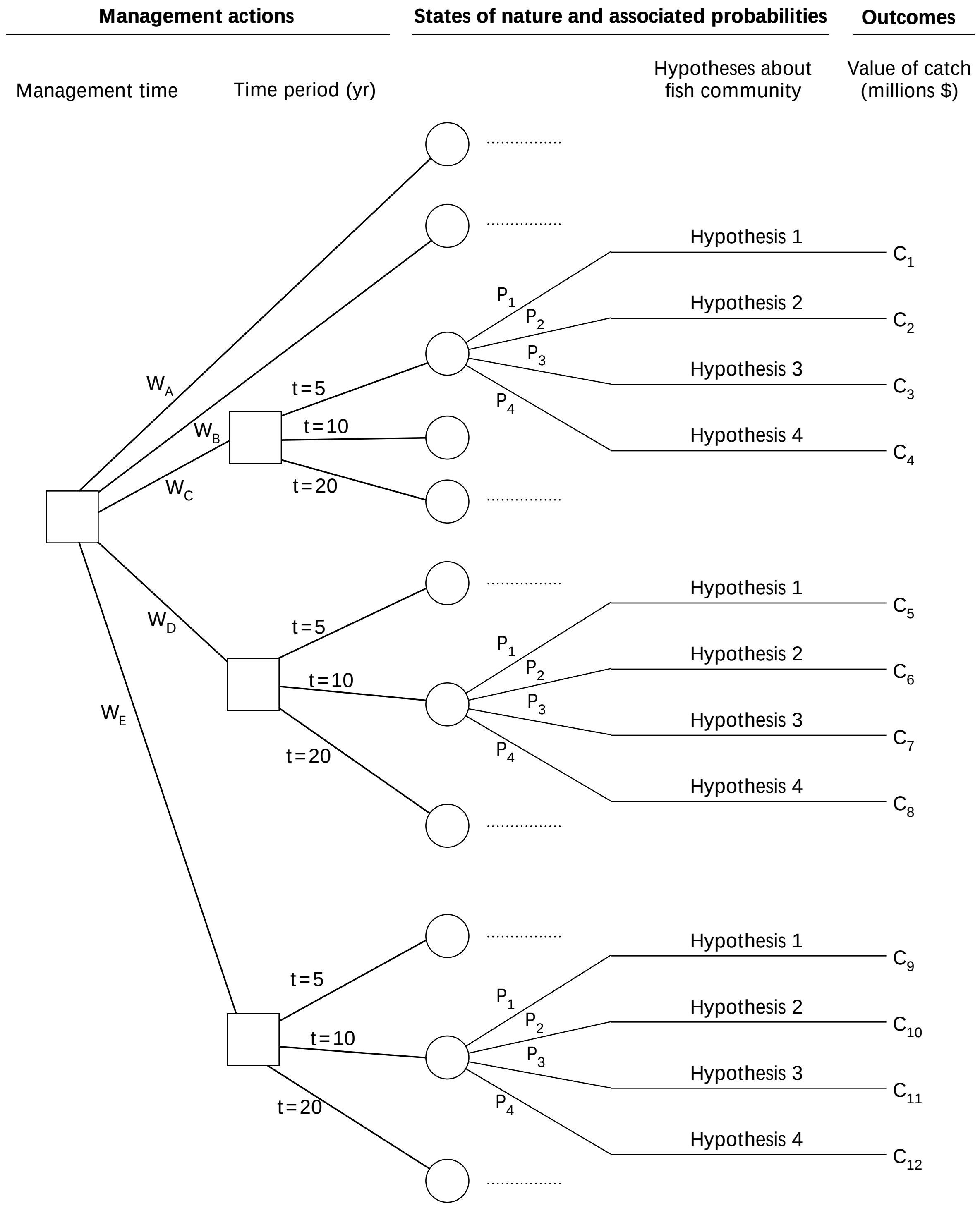


FIGURE 8.7 *Decision tree for the analysis of various management actions in Sainsbury's (1988) large-scale fishing experiment. Management strategies (W_A to W_E), time periods, hypotheses, and outcomes are described in the text and Table 8.3. Only a subset of the complex branches is shown.*

TABLE 8.3 *Results of Sainsbury's (1991) calculations of the benefits of different designs for an active adaptive management experiment on groundfish in Australia. Management strategies W_A to W_E are defined in the text; they differed in how much fishing occurred and when, what type of gear was used, and whether the strategies were based only on existing information as of 1985 (W_A and W_B) or on information resulting from the active adaptive experiment (W_C, W_D, W_E). W_B,$_1$ to W_B,$_4$ refer to four different long-term harvesting strategies; time period, t, is the duration of the experiment in years. Expected values of catch are in millions of Australian dollars. See text for details. (Adapted from Sainsbury 1991.)*

Strategy	Expected Value of catch (millions $)
W_A	9.96
$W_{B,1}$	27.2
$W_{B,2}$	35.4
$W_{B,3}$	31.8
$W_{B,4}$	9.96
W_C,t = 5	35.6
W_C,t = 10	29.7
W_C,t = 20	21.2
W_D,t = 5	37.4
W_D,t = 10	37.2
W_D,t = 20	36.3
W_E,t = 5	40.6
W_E,t = 10	40.5
W_E,t = 20	38.6

Bayesian statistical analysis. Several management strategies, implemented from 5 to 20 years, were simulated under various scenarios, starting in 1985. After this simulated learning period, the model then determined which of the four hypotheses was the most likely, which in turn would suggest a long-term management plan. Thus, Sainsbury's expected value of the catch included the value during both the learning period and the subsequent period of implementing the derived optimal action.

Sainsbury (1991) found that the expected value of the catch was maximized at AUS $40.6 million by applying strategy W_E for 5 years. This experimental strategy had some replicate areas open to trawling and others closed. The other adaptive management regimes had lower expected values (Table 8.3), which illustrates the benefits of applying decision analysis to compare different designs of experimental management plans. Without this type of rigorous analysis, a suboptimal experimental design might have been chosen. The value of collecting information was included in the calculated economic value of the catch because a management strategy that produced high-quality data during the learning period led to improved understanding of which of the four hypotheses was responsible for the decline in abundance of the valuable species. This approach allowed a more accurate decision to be made about which long-term harvesting strategy was most likely to reverse the problem and increase the value of the catch. (Incidentally, Sainsbury et al. 1997 reported that the experimental management strategy W_E generated data by 1991 that strongly supported the fourth hypothesis—that trawling detrimentally affected the habitat of the more valuable groundfish species. Trawling was subsequently reduced.)

8.7 Value of Information

By taking uncertainty into account quantitatively in decision analyses, analysts can quantify the effects of considering or reducing uncertainties when making decisions. Several types of analyses are possible: expected value of including uncertainty (EVIU), expected value of sample information (EVSI), expected value of perfect information (EVPI), and expected value of experimental or adaptive management. (See Morgan and Henrion 1990 for more details.)

The expected value of including uncertainty (EVIU) provides a measure of how much better the expected value of some decision will be if analysts consider uncertainty through a decision analysis, as opposed to the common approach of using only the best point estimates of parameters to make decisions. EVIU is calculated as the difference between the expected outcome of a decision based on a probabilistic decision analysis (the Bayes decision) and a decision based only on using the best point estimates of uncertain parameters (the deterministic decision). Therefore, EVIU represents the increase in expected benefits or reduction in expected losses that results from using decision analysis and can be used to determine whether to spend the additional time necessary to gather the data and complete a decision analysis. Note that EVIU is always ≥ 0 because the Bayes decision accounts for the potentially useful information (contained in the uncertainties) that is lost if these uncertainties are ignored.

Another measure, the expected value of sample information (EVSI), estimates the benefits of reducing uncertainties by collecting additional data through a monitoring program or adaptive management experiment. EVSI requires calculating the expected value of a decision made with improved information, compared to the current level of uncertainty. In practice, this value is estimated by adjusting the probabilities placed on the states of nature to reflect more certainty (i.e., making the probability distribution more precise or more accurate) and then repeating the decision analysis using this adjusted distribution. EVSI is then the difference between this value and the expected value of the Bayes decision. The effect of additional information on the probabilities for the states of nature can sometimes be estimated using sampling theory. If the cost of carrying out the sampling program can be estimated, the ratio of the benefits (EVSI) to the costs provides one way to evaluate the effectiveness and efficiency of planned sampling programs and to allocate research budgets among different sampling programs.

The expected value of perfect information (EVPI) is a measure of the increase in expected benefits or decrease in expected losses if we could forecast the outcomes of management actions with complete certainty. Consequently, this value is the maximum amount that we should be willing to pay for research that will generate information and reduce uncertainties. Although this value is hypothetical because uncertainties are always present, it is often instructive to compute it because this value provides an upper bound on EVIU and EVSI.

These general concepts of value of information are directly relevant to adaptive or experimental management because the expected value of an experiment can be calculated explicitly using decision analysis. For instance, Table 8.3 shows Sainsbury's (1991) estimates of the expected value of various management plans as calculated before the experiment began in 1985. For example, the expected value of the catch from allowing the existing trawl fishery to continue (Strategy $W_{B,4}$) was $9.96 million, given the uncertainty that existed in 1985. Immediate implementation of long-term harvesting strategy $W_{B,2}$ (moderate-intensity trap fishery) in 1985, without collecting any additional information, would have increased the expected value of the catch to $35.4 million. This maximum expected value of the catch represents what could have been realized given the level of uncertainty in 1985. However, several of the proposed experimental management strategies (W_C, W_D, and W_E) produced even larger expected values of catch (Table 8.3) because of the value of the information about the uncertain biological hypotheses that were generated by the experiment. For example, as noted previously, the experimental strategy W_E with a learning period of 5 years maximized the expected value of the catch at $40.6 million. This amount was $5.2 million more than the next best strategy that could have been implemented in 1985 without doing experimental management ($W_{B,2}$).

8.8 Quantifying Management Objectives

For a decision analysis to rank the management options, one or more management objectives must be identified. However, disagreement about what the management objective should be is common, as in the land use issue in the Tatshenshini described previously (McDaniels 1992). Disagreement often occurs when management objectives are based on consultation with a wide range of managers and stakeholders (Bell et al. [editors] 1977). These disagreements can be resolved by repeating the decision analysis using several different management objectives, each representing a different viewpoint. The key question in such analyses is how much the optimal decision changes when different management objectives are used. As noted previously in the sensitivity analysis section, addressing this issue can help resolve conflicts by identifying which assumptions or elements of the objectives lead to different recommended management actions. In some cases, participants may not disagree once the quantitative decision analysis is done.

Conflicting management objectives can be treated more formally using multi-attribute utility theory (Keeney and Raiffa 1976). Utility is a unitless measure of satisfaction obtained from different quantitative outcomes of decisions. Utility analysis converts into common units (utilities) different kinds of outcomes, or attributes, such as dollar value of timber and an index of biodiversity. Utility functions permit this conversion and the shapes of these functions reflect the degree of risk aversion of the stakeholder. Once converted to utilities, these attributes can be combined into a weighted average utility, where weightings placed on different attributes reflect their relative importance to different interest groups.

Multi-attribute utility analysis thus provides a quantitative method for incorporating multiple and conflicting management objectives into the decision-making process. The disadvantage of combining these objectives into a single weighted average utility is that the trade-offs implicit in combining multiple objectives are hidden from the decision-maker. For that reason, it is often preferable to show the separate attributes as functions of the management decision along with the results of the multi-attribute utility analysis. This explicitly shows decision-makers the trade-offs that are inherent in particular decisions.

8.9 Communicating Uncertainty and Results of Decision Analyses

To establish confidence in the analysis, all users of the results of a decision analysis must be informed not only of the optimal decision, but also of the assumptions for which that action is optimal. This approach is necessary because, as noted under sensitivity analysis, different parameter values, model structures, or management objectives can sometimes lead to a different optimal decision. One of the advantages of decision analysis is that these assumptions are made explicit, and consequently the effects of these assumptions on the optimal decision can be explored quantitatively. However, these advantages are lost unless the decision analyst communicates these results clearly and effectively to decision-makers. Several steps can be taken to ensure good communication.

First, the decision-making process used to identify the optimal decision must be adequately documented. This information includes full documentation of the data, key assumptions, and methods; a list of which factors were considered uncertain and why; the results and implications of sensitivity analyses for the decision-maker; and limitations of the decision analysis. Such documentation shows the decision-maker exactly how an optimal decision was derived and provides an indication of how robust that decision is to the assumptions. The documentation also allows decisions to be made consistently when the same decision must be made by different people at a different time or place. As well, good documentation allows decisions to be subjected to an iterative process of external peer review and evaluation.
In doing so, analysts and decision-makers can learn from successes and failures, leading to progressive improvements in decision-making.

The second step toward good communication is to choose appropriate methods for presenting results of decision analyses and sensitivity analyses. What seems most appropriate for technical analysts may be confusing for non-specialists. For instance, Ibrekk and Morgan (1987) showed that the conventional way for scientists to express uncertainty in some quantity, a probability distribution, is not the most effective way to convey understanding of that uncertainty to others (i.e., they found a cumulative probability distribution to be best). In some cases, a computer-based hierarchical information system has been used to present and explain results of analyses (e.g., Gobas 1993). Such information systems allow the user to select the level of detail, ranging from summary graphs to complete, detailed numerical tables.

A final measure to ensure good communication between various individuals is to involve decision-makers and other interested parties in the analysis from the beginning. This step can be done through workshops and by allowing users to conduct "what-if" runs with simulation models. Early participation by decision-makers and stakeholders avoids misunderstandings and misinterpretations of key assumptions, data, methods, and results. It also increases the chance that these important groups will accept the results and will support the analysis by providing necessary information.

8.10 Benefits and Limitations of Decision Analysis

8.10.1 Benefits

To summarize, several key benefits of decision analysis have made it an increasingly popular method for quantitatively considering uncertainties when making decisions in resource management.

1. By systematically accounting for complexities and uncertainties, decision analysis can improve the quality of decisions made, resulting in more favourable outcomes over the long term.
2. Using a decision tree to structure a problem helps identify what specific data and assumptions are needed to perform the analysis.
3. Going through a formal decision analysis requires explicit statements of the assumptions, parameter values, and management objectives, including views about risk aversion.
4. Decision analysis allows explicit comparison and ranking of alternative management actions.
5. Sensitivity analyses help to set priorities for future research and establish confidence in the analysis by identifying the robustness of a recommended action.
6. Explicitly taking uncertainties into account permits calculation of the benefits of considering uncertainty compared to only using the best point estimates (EVIU), and the value of reducing these uncertainties through the implementation of a research sampling program (EVSI).
7. A systematic approach documents the method by which decisions were reached and thus indicates which methods of analysis and management actions work and why.
8. Decision analysis can be used for conflict resolution between interest groups.

8.10.2 Limitations

As with any method that assists decision-making, decision analysis also has its limitations. A major limitation is that the amount of data required to conduct a decision analysis of a complex problem can be large. States of nature, probabilities on those states, and outcomes of management actions must all be quantified to apply decision analysis. In many cases, these data may not be available in sufficient quantity or quality to allow formal decision analysis. Another limitation of decision analysis is that quantifying management objectives is sometimes difficult. This situation is especially problematic when diverse user groups or stakeholders are part of the decision-making body or are involved in consultations with managers. Under these circumstances, identifying quantitative indicators of management objectives can be difficult even if multiattribute utility analysis is applied.

These limitations can, in some cases, make application of decision analysis impossible or unwarranted. When this happens, management actions should be taken cautiously, given the inevitable presence of uncertainties.

Another limitation of decision analysis stems not so much from the method itself as from the way in which results are used. As described previously, decision theorists define "risk" as "expected loss." Thus, when decision analysis compares the expected values of outcomes for various possible management actions, it essentially calculates the risk associated with each action. However, when managers or scientists present such results to stakeholders, they may interpret them quite differently. Substantial research shows that such people often perceive risks quite differently from the amount of risk estimated from quantitative analyses (Slovic 1987). The magnitude of this difference depends on factors such as the amount of control over the risk, the level of trust in the experts, and the immediacy of the effect (Slovic 1987).

8.10.3 Evaluation of quality of decisions

The benefits and limitations of decision analysis lead to two important points about how decisions should be evaluated. First, *the quality of decisions should be evaluated based on the process used to make them, not on their short-term outcome.* This view is based on the observation that, because of the complexity of forest ecosystems and the number of factors influencing them, favourable outcomes might arise as much from fortuitous events as from good decisions. Thus, for instance, if chance events happened to lead to a good outcome in some situation, in spite of an incorrect decision, managers might unjustifiably conclude that the decision they made was correct and they might repeat it in similar circumstances. However, the chance events might not occur again in the managers' favour. Similarly, a correct decision might lead to some detrimental effect because of an unfavourable chance event that coincided with the decision. For this reason, conclusions about the quality of a decision should *not* be based on short-term *outcomes*; they should be based on how systematically, rigorously, and comprehensively the decision options were evaluated before making the decision. Studies show that decisions based on rigorous analyses that quantitatively account for uncertainties will, *in the long term*, produce better results than decisions made using other approaches (Von Winterfeldt and Edwards 1986). Thus, decisions that are based on a rigorous approach to analyzing options and uncertainties should be labelled "good" decisions, whereas others should be described as unacceptable, *regardless of the short-term outcomes of any particular decision.*

The second point is that the decision-making process should be judged not on an absolute scale, but relative to other methods available. Decision analysis has some potentially serious limitations, but few alternative methods have been demonstrated to provide a better approach to using information on the uncertainties and complexities inherent in resource management decisions. Because of this, decision analysis is being increasingly applied to a wide range of problems in toxicology, water quality, forestry, fisheries, and wildlife management. However, methods for quantitative policy analysis are continually being improved, and analysts should be aware of developments in the field so that they use the best methods available to make decisions.

8.11 Final Recommendations for Decision Analysts and Decision-makers

- Do not push scientists to "state their best estimate despite the uncertainties" because this effectively ignores uncertainties and will often lead to management actions with undesirable results. Instead, for choosing among options, use a systematic method such as decision analysis, which takes uncertainties into account explicitly.
- Do not forget the caveats and limitations of the various components of a decision analysis. For example, recognize the trade-offs between the complexity of models and their reliability. Acknowledge the assumptions behind formulation and parameterization of models, and use sensitivity analyses to explore how these factors affect the optimal decision.
- When doing a decision analysis, adhere to the following guidelines to ensure that the decision-making process is the best available:
  1. Clearly identify the main goal of the decision analysis.
  2. Ensure that interaction occurs early and periodically among scientists, analysts, decision-makers, the public, user groups, and other stakeholders.
  3. Document all steps in the analysis.
  4. Do not assume that everyone will agree with your methods (e.g., Bayesian statistics), estimates of parameters, or interpretation of data.
  5. State the assumptions and data used, carefully qualify the conclusions, and clearly define the limits of the analysis.
  6. Present extensive sensitivity analyses that focus on:

     a) how the rank order of management options changes with different assumptions; and

     b) research priorities—the most important areas for getting more data.
  7. Be cautious: not only could the analysis be incomplete but it will almost certainly be missing components. These factors may affect your results.
  8. When communicating information about risks or uncertainties, think about what it is like *not* to know the material.
  9. The entire process of decision analysis and communication should be iterative, continually moving toward improving decisions.
  10. Insist on objective science and rigorous external peer review of analyses.
  11. When decision analysis is used to evaluate different proposed adaptive management actions, implement a monitoring program in conjunction with the chosen action to ensure that the maximum amount of information is obtained.
  12. Recognize that decision analysis is only one part of the whole process for making decisions—it is *NOT* the entire process. However, if decision analysis is one component, it can help improve environmental decision-making.

- Not all circumstances warrant a full, formal quantitative decision analysis—justifiable usage of decision analysis is case-specific. For example, decision analysis is more feasible if at least some reliable data are available and clear management objectives are stated. Furthermore, decision analysis is more appropriate when costs of incorrect decisions are potentially large. First-time users of this approach are encouraged to use the references here and to discuss the approach with experts who have previously used decision analysis. Regardless of the specific situation, it is always worth at least thinking about a decision-making problem in terms of the components of decision analysis as described, even if final calculations are never carried out due to limitations in data or other problems. The mere process of describing each component helps to clarify and organize the decision-making process and to identify research needs.

Acknowledgements

We are grateful to Vera Sit, Brenda Taylor, Darrell Errico, and Milo Adkison for many useful suggestions, comments, and discussions. Milo Adkison also suggested the format for Table 8.1. Several reviewers also provided helpful comments: Russ Horton, Wendy Bergerud, Michael Stoehr, Jeff Stone, and Peter Ott.

References

Adkison, M.D. and R.M. Peterman. 1996. Results of Bayesian methods depend on details of implementation: an example of estimating salmon escapement goals. Fish. Res. 25:155–70.

Bell, D.E., R.L. Keeney, and H. Raiffa (editors). 1977. Conflicting objectives in decisions. J. Wiley, New York, N.Y.

Berger, J.O. and D.A. Berry. 1988. Statistical analysis and the illusion of objectivity. Am. Sci. 76:159–65.

Bergerud, W.A. and W.J. Reed. [n.d.]. Bayesian statistical methods. This volume.

Bergh, M.O. and D.S. Butterworth. 1987. Towards rational harvesting of the South African anchovy considering survey imprecision and recruitment variability. S. African J. Marine Sci. 5:937–51.

Box, G.E.P. and G.C. Tiao. 1973. Bayesian inference in statistical analysis. Addison-Wesley, Reading, Mass.

Clark, C.W. 1990. Mathematical bioeconomics: the optimal management of renewable resources. J. Wiley, New York, N.Y.

Clemen, R.T. 1996. Making hard decisions: an introduction to decision analysis. 2nd ed. Duxbury Press, Wadsworth Publ. Co., Belmont, Calif.

Cohan, D., S.M. Haas, D.L. Radloff, and R.F. Yancik. 1984. Using fire in forest management: decision making under uncertainty. Interfaces 14:8–19.

Crome, F.H.J., M.R. Thomas, and L.A. Moore. 1996. A novel Bayesian approach to assessing impacts of rain forest logging. Ecol. Applic. 6:1104–23.

Edwards, A.W.F. 1992. Likelihood: expanded edition. Johns Hopkins Univ. Press, Baltimore, Md.

Ellison, A.M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecol. Applic. 6:1036–46.

Errico, D. 1989. Choosing stand density when spacing lodgepole pine in the presence of risk of pest attack. B.C. Min. For., Victoria, B.C. Res. Rep. 89004-HQ.

Frederick, S.W. and R.M. Peterman. 1995. Choosing fisheries harvest policies: when does uncertainty matter? Can. J. Fish. Aquat. Sci. 52:291–306.

Gobas, F.A.P.C. 1993. A model for predicting the bioaccumulation of hydrophobic organic chemicals in aquatic food-webs: application to Lake Ontario. Ecol. Modeling 69:1–17.

Hairston, N.G. (editor). 1989. Ecological experiments: purpose, design, and execution. Cambridge Univ. Press, Cambridge, U.K.

Henrion, M. and B. Fischhoff. 1986. Assessing uncertainty in physical constants. Am. J. Physics 54:791–7.

Hilborn, R., E.K. Pikitch and M.K. McAllister. 1994. A Bayesian estimation and decision analysis for an age-structured model using biomass survey data. Fish. Res. 19:17–30.

Holbert, D. and J.C. Johnson. 1989. Using prior information in fisheries management: A comparison of classical and Bayesian methods for estimating population parameters. Coastal Manage. 17:333–47.

Howard, R.A. 1988. Decision analysis: practice and promise. Manage. Sci. 34:679–95.

Ibrekk, H. and M.G. Morgan. 1987. Graphical communication of uncertain quantities to non-technical people. Risk Analysis 7:519–29.

Keeney, R.L. 1982. Decision analysis: an overview. Operations Res. 30:803–38.

______. 1992. Value-focused thinking. Harvard Univ. Press, Cambridge, Mass.

Keeney, R.L. and H. Raiffa. 1976. Decisions with multiple objectives: preferences and value trade-offs. J. Wiley, New York, N.Y.

Lindley, D.V. 1985. Making decisions. Wiley Interscience, New York, N.Y.

Lord, G.E. 1976. Decision theory applied to the simulated data acquisition and management of a salmon fishery. Fish. Bull. (U.S.) 74:837–46.

McAllister, M.K. and R.M Peterman. 1992a. Decision analysis of a large-scale fishing experiment designed to test for a genetic effect of size-selective fishing on British Columbia pink salmon (*Oncorhynchus gorbuscha*). Can. J. Fish. and Aquat. Sci. 49:1305–14.

______. 1992b. Experimental design in the management of fisheries: a review. N. Am. J. Fish. Manage. 12:1–18.

McDaniels, T. 1992. A multiple objective decision analysis of land use for the Tatshenshini-Alsek area. Appendix to report for B.C. Commission on Resources and Environment, Victoria, B.C.

Maguire, L.A. 1986. Using decision analysis to manage endangered species populations. J. Environ. Manage. 22:345–60.

Maguire, L.A. and L.G. Boiney. 1994. Resolving environmental disputes: a framework incorporating decision analysis and dispute resolution techniques. J. Environ. Manage. 42:31–8.

Mapstone, B.D. 1995. Scalable decision rules for environmental impact studies: effect size, Type I, and Type II errors. Ecol. Applic. 5:401–10.

Morgan, M.G. and M. Henrion. 1990. Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge Univ. Press, Cambridge, U.K.

Nyberg, J.B. [n.d.]. Statistics and the practice of adaptive management. This volume.

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, K.E. Abu-Saba, and A.R. Flegal. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. Ecol. Applic. 4:16–30.

Parkhurst, D.F. 1984. Decision analysis for toxic waste releases. J. Environ. Manage. 18:105–30.

Parma, A.M. and R.B. Deriso. 1990. Experimental harvesting of cyclic stocks in the face of alternative recruitment hypotheses. Can. J. Fish. Aquat. Sci. 47:595–610.

Pearse, P.H. 1976. Timber rights and forest policy in British Columbia. Rep. Royal Comm. For. Res., Victoria, B.C.

Peel, A.L. 1991. Forest resources commission: the future of our forests. B.C. Forest Resources Commission, B.C. Min. of For., Victoria, B.C.

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

______. 1990b. The importance of reporting statistical power: the forest decline and acidic deposition example. Ecology 71:2024–7.

Peterman, R.M. and M.K. McAllister. 1993. A brief overview of the experimental approach to reducing uncertainty in fisheries management–an extended abstract. *In* Risk evaluation and biological reference points for fisheries management. S.J. Smith, J.J. Hunt, and D. Rivard (editors). Can. Spec. Pub. Fish. Aquat. Sci. 120:419–22.

Press, S. J. 1989. Bayesian statistics: principles, models, and applications. J. Wiley, New York, N.Y.

Raiffa, H. 1968. Decision analysis: introductory lectures on choices under uncertainty. Addison-Wesley, Don Mills, Ont.

Reckhow, K.H. 1994. Importance of scientific uncertainty in decision making. Environ. Manage. 18:161–6.

Render, B. and R.M. Stair. 1988. Quantitative analysis for management. Allyn and Bacon, Boston, Mass.

Sainsbury, K.J. 1988. The ecological basis of multispecies fisheries, and management of a demersal fishery in tropical Australia. *In* Fish Population Dynamics, J.A. Gulland (editor), 2nd ed. J. Wiley, New York, N.Y. pp. 349–82.

______. 1991. Application of an experimental approach to management of a tropical multispecies fishery with highly uncertain dynamics. ICES (International Council for the Exploration of the Sea) Marine Sci. Symp. 193:301–20.

Sainsbury, K.J., R.A. Campbell, R. Lindholm, and A.W. Whitelaw. 1997. Experimental management of an Australian multispecies fishery: examining the possibility of trawl induced habitat modification. *In* Global trends: fisheries management. E.L. Pikitch, D.D. Huppert, and M.P. Sissenwine (editors). American Fisheries Society Symposium, Vol. 20, Bethesda, Md, pp. 107–12.

Scientific Panel for Sustainable Forest Practices in Clayoquot Sound. 1995. Sustainable ecosystem management in Clayoquot Sound: planning and practices. Rep. 5, Victoria, B.C.

Simberloff, D.S. and L.G. Abele. 1976. Island biogeographic theory and conservation practice. Science 191:285–6.

Slovic, P. 1987. Perception of risk. Science 236:280–5.

Sokal, R. and F.J. Rohlf. 1969. Biometry: the principles and practice of statistics in biological research. W.H. Freeman, San Francisco, Calif.

Stahl, G., D. Carlsson, and L. Bondesson. 1994. A method to determine optimal stand data acquisition policies. For. Sci. 40:630–49.

Thompson, G.G. 1992. A Bayesian approach to management advice when stock-recruitment parameters are uncertain. Fish. Bull. (U.S.). 90:561–73.

Von Winterfeldt, D. and W. Edwards. 1986. Decision analysis and behavioral research. Cambridge Univ. Press, Cambridge, U.K.

Walters, C.J. 1981. Optimum escapements in the face of alternative recruitment hypotheses. Can. J. Fish. Aquat. Sci. 38:678–89.

______. 1986. Adaptive management of renewable resources. MacMillan, New York, N.Y.

Westland Resource Group. 1995. Environmental indicators for land and resource management planning. Prep. for B.C. Min. of Environ., Lands and Parks, Victoria, B.C.

9 SELECTING APPROPRIATE STATISTICAL PROCEDURES AND ASKING THE RIGHT QUESTIONS: A SYNTHESIS

BRUCE G. MARCOT

Abstract

In this chapter, I synthesize statistical approaches and considerations for adaptive management studies. I review approaches to learning from management actions and address questions of space and time. I also present a set of guidelines for asking the right questions about statistical reliability, for selecting the appropriate adaptive management study, and for guiding how different types of information can contribute at different stages in adaptive management. These guidelines are presented in a table, which can be used as a decision tree to determine the best kinds of studies for each step in the adaptive management process, and the most appropriate use of exisiting information.

9.1 Introduction

How should managers and researchers select an approach for designing an adaptive management study and analyzing the results? The chapters in this report provide some guidance; for example, Nemec (this volume, Chap. 2), summarizes principles of experimental design, and Schwarz (this volume, Chap. 3) lists types of nonexperimental and experimental designs. Other publications (e.g., Green 1979), while not specific to adaptive management as defined in this volume, also provide guidance on designing ecological studies. This chapter reviews issues to consider in designing adaptive management studies, synthesizes the methods discussed in preceding chapters of this report, and summarizes the roles different types of information can play in adaptive management.

Statistical approaches and study designs can be selected only when the management question is first well articulated. In the first section of this chapter, I review three types of monitoring, differentiated by the types of question they each address, and then address how the spatial and temporal elements of a management question can influence study design. In the second section, I review the characteristics of powerful studies and the principles of experimental design. The third section summarizes various types of information (including existing data, retrospective studies, and nonexperimental studies) and experimental studies, and how they can contribute to adaptive management. In the final section, I discuss some points to consider in interpreting and communicating the results from adaptive management studies, and in particular the difficulty in "unravelling the causal web." Throughout this chapter, I use the oversimplistic labels "researcher" and "manager," fully realizing that in the real world many resource professionals don both hats.

9.2 Types of Questions Addressed in Adaptive Management

A little experience often upsets a lot of theory.
– Cadman

The B.C. Ministry of Forests defines adaptive management as a formal process entailing problem assessment, study design, implementation, monitoring, evaluation, and feedback (B.C. Ministry of Forests 1996). In this approach, management activities are crafted as experiments to fill critical gaps in knowledge. The key questions are: (1) To what extent did the managment action lead to the measured outcome? and (2) Are our assumptions valid about how the system works?

Other institutions use the term "adaptive management" differently. For example, the USDA Forest Service incorporates the general concepts of adaptive management into its planning, but not as a formal process. Regardless of the definition of adaptive management and how it is institutionalized, monitoring activities and evaluation of data are key steps in adaptive management. The statistical approaches discussed in this report can help in both the design of monitoring activities and in the interpretation of data.

9.2.1 Types of monitoring

There are three types of monitoring: implementation monitoring, effectiveness monitoring, and validation monitoring. Each type of monitoring serves a unique function in an adaptive management study.

Implementation monitoring

Implementation monitoring (or compliance monitoring) essentially asks: Have the management guidelines been implemented correctly (Collopy et al. 1993)?

Correct implementation can be determined by a complete census of all activities or by sampling activities stratified by administrative unit or location. Obviously, asking more detailed questions of the effects and validity of particular management activities should proceed only when they have been correctly implemented. Implementation monitoring, however, does not teach us about effects of management actions. Thus, the focus of adaptive management is effectiveness and validation monitoring.

Effectiveness monitoring

Effectiveness monitoring asks: Are the management guidelines and activities producing the desired effects? Do the management activities really alter the biophysical conditions as expected? Many questions can be asked of the effects of management guidelines. Highest priority should be directed to potential effects that have the most serious economic, biological, or ecological ramifications, and those carrying the greatest uncertainty.

Validation monitoring

Validation monitoring, technically the most difficult of the three kinds of monitoring, asks: Are the ultimate expectations for the guidelines being met? Are the basic assumptions about how the biophysical system operates really correct, or does it operate in a very different way that would invalidate the selected management approach? If so, how?

Validation monitoring may be used to validate ecosystem models (Gentiol and Blake 1981), which is vital to ensuring the models' successful and appropriate use. In adaptive management, validation monitoring should focus on the ecosystem elements that have the greatest implications on the decision about the best course of action. Problem assessment—identifying which relationships to validate—is the first step of adaptive management.

9.2.2 Issues of space and time

Issues of space and time will in part determine the type of study design that is possible. For example, studies of large geographic areas may preclude replication, suggesting before-after-control-impact paired (BACI-P) study (Schwarz, this volume, Chap. 3). Similarly, long response times may suggest retrospective analysis of past actions to provide a preliminary assessment of the impact of a proposed action.

Issues of space

The five kinds of spatial effects to consider can influence the design of a study as well as the interpretation of its results.

1. What is the influence of on-site management activities on off-site conditions? That is, local management may influence remote conditions, both directly and indirectly (Loehle 1990). An example is the downstream effect of stream temperature or sedimentation on fish populations due to local reduction, removal, or restoration of riparian vegetation cover.

2. What is the relative influence of off-site management activities on on-site (desired) conditions? On-site conditions can be influenced by other off-site activities. For example, despite protection of old-growth forest groves, some arboreal lichens might nonetheless decline because of degraded air quality from industrial pollutants originating elsewhere in the airshed. The potential influence of downstream dams and fish harvesting on the abundance of local fish populations is another example.

3. To what *degree* do local management activities influence the on-site (desired) conditions? That is, to what extent do background noise and other environmental factors affect on-site conditions? Local management may influence only a portion of the total variation in local conditions. For example, providing local breeding habitat only partially succeeds in conserving populations of neotropical migratory birds, whose numbers may still decline due to pesticide loads or habitat loss encountered during wintering in the neotropics.

4. What is the relative influence of conditions and activities from different spatial scales, particularly the effects on local stand-level conditions from broader landscape-level factors? That is, desired conditions and management actions are best addressed at appropriate scales of geography. As examples, effects of forest management on abundance of coarse woody debris are best assessed at the stand level; effects of forest management on vegetation conditions that affect visual quality or goshawk (*Accipiter gentilis*) habitat are best assessed at the landscape level; and effects of overall management policy and ownership patterns on grizzly bear (*Ursus arctos*) populations are best assessed at subregional or regional levels.

5. What are the cumulative effects of stand-level treatments as they spread across the landscape? For example, wind fetch and thus wind speed may increase as clearcuts become wider with sequential, adjacent cuts. Thus, the windthrow hazard in one cutblock may increase as adjacent areas are cut, and the windthrow hazard in those cutblocks cannot simply be extrapolated from the hazard measured in a single cutblock surrounded by trees.

For each of these five kinds of spatial effects, adaptive management monitoring studies would be designed and implemented differently. Where this is not possible, spatial influences should at least be acknowledged as potential sources of variation and included in the analysis.

Issues of time

Answering questions about time effects can help distinguish true cause from non-causal correlation, and treatment effects from natural variation. Three typical time scale issues follow.

1. What are the response times of variables? For some variables, response may be apparent in a relatively short period of time; others may respond more slowly. Examples are the relatively short and quick response time of seedling survival compared with the long and slow response times associated with many biodiversity indices (e.g., changes in grizzly bear populations).
2. What are the lag times of variables? Some variables may not immediately respond to a treatment or may depend greatly on site history. For example, because acorn woodpeckers (*Melanerpes formicivorous*) show high fidelity to particular sites, a lag will exist before they respond to the removal of granary trees (Ligon and Stacey 1996). This lack of short-term response should not lead one to conclude that management actions—in this example, the reduction or removal of granary trees—have no effect. Sometimes these lags in response result when conditions from prior time periods overwhelm or influence responses from current actions. For example, the intensity of a fire will be influenced by site history, in addition to current management actions. Thus short-term changes in a response variable may reflect both the management action and past site history. Some time-lag effects can be quite variable and manifest as non-monotonic (up and down) trends over the long term. For example, annual non-monotonic variations in bird populations—both increases and decreases—may belie truer long-term declines in some population counts (Thomas and Martin 1996).
3. What are the cumulative effects of a variable over time? Some variables do not make a mark except over time or until a particular threshold has been exceeded. An example is the adverse effect of certain pesticides on wildlife reproduction. The detrimental effect may not be apparent until the pesticide concentrations reach a particular level of toxicity (Tiebout and Brugger 1995).

The design of adaptive management studies and selection of analysis methods are guided in part by these considerations of space and time. For example, replication is one major consideration in designing studies. Given a large geographic area, as tends to be the focus in ecosystem management, or a rare condition, such as a threatened species population, are spatial replicates possible? That is, can landscapes or threatened populations be replicated at all, or in adequate numbers? If the conditions cannot be replicated, then pseudoreplication (e.g., dividing a single area into smaller blocks) may be the only recourse (Hurlbert 1984). Alternatively, other kinds of studies (e.g., analytical surveys, expert testimony) might help in assessing the impact of the treatments, although they do not allow strong inference about cause. Similarly, long response times and time lags make temporal replication difficult. Retrospective studies (see Smith, this volume, Chap. 4) provide one alternative for gaining insight into the long-term effects of management actions. In cases where either spatial or temporal replication is severely limited, a higher probability of Type I and II errors might need to be tolerated (see Anderson, this volume, Chap. 6).

In some cases, a powerful adaptive management study may be possible but managers, decision-makers, industries, or other interested bodies may not be willing to bear the cost, duration, and tight controls on management activities. The consequences of not using an optimum study must be explicitly considered and understood by all.

9.3 Considerations in Designing an Adaptive Management Study

9.3.1 Characteristics of a powerful adaptive management study

To help in evaluating management actions and validating functional and causal relationships, an adaptive management study should be consistent (i.e., should represent the system of interest), accurate, precise, and unbiased (see Routledge, this volume, Chap. 5). Managers and researchers should work together in designing an adaptive management study that represents the real system and provides information within acceptable limits of Type I and Type II errors (Anderson, this volume, Chap. 6). They may also want to consider the trade-offs inherent in relaxing any of the conditions, such as accepting a lower but still acceptable level of precision in exchange for lower cost or more rapid results. The study design should also be independently reviewed to assess its capability to meet the desired (and often conflicting) criteria of high consistency, high accuracy, high precision, and low bias.

9.3.2 What managers need to ask of reliability

Managers should ask four general questions regarding the reliability of adaptive management studies and their results.

1. What *confidence* can I have in the results of this adaptive management study, particularly for avoiding false positives? Statistically, this question can be answered by calculating the probability of a Type I error (Anderson, this volume, Chap. 6).
2. What *power* do the results provide for avoiding false negatives (Anderson, this volume, Chap. 6)? Statistically, this can be answered by calculating the probability of a Type II error (although Bayesian approaches differ significantly in not dealing with questions of confidence and power). Type I and Type II errors hold different implications for managers (Marcot 1986; Anderson, this volume, Chap. 6). For example, if the adaptive management study is aimed at determining adverse effects of some management activity on a wildlife species that is threatened, then managers may be more tolerant of a Type I error than of a Type II error. However, if the species is not threatened and the activity results in important commodity production and economic return, then they may be more tolerant of a Type II error.
3. What is the *relevance* of the results? How representative is the study of other sites or conditions? Some studies may reveal only local conditions and the chance effects of unique site histories, rather than overall effects, or they may pertain to only one vegetation type or climatic condition. The manager should know the contexts under which results apply. For example, results of a forest thinning operation may apply to only a particular initial stand density or forest type.
4. Were the effects truly a result of the *management activity*? This question cuts to the heart of separating cause from noise, and determining what really influenced the outcome. The experimental studies that are central to adaptive management are designed to determine causality. Researchers and managers should not assume that demonstration of pattern and correlation constitutes valid evidence of causation.

9.3.3 Principles of experimental design

To help ensure success in evaluating management actions, researchers should review adaptive management studies for the four main principles of experimentation: randomization, replication, blocking, and representation (see Nemec, this volume, Chap. 2). *Randomization* reduces bias. *Replication* allows an estimation of variance, which is vital for confirming observed differences. *Blocking* increases precision and reduces cost and sample size. *Representation* helps to ensure study of the correct universe of interest.

In the real world, these four principles cannot always be met and compromises are necessary. It is often impossible to fully randomly allocate treatments, such as forest clearcuts or fire locations. In such cases, study sites may be randomly selected from existing clearcuts or fire locations, resulting in nonexperimental studies (e.g., observational studies, analytical surveys, retrospective studies, or impact studies; see Schwarz, this volume, Chap. 3). When interpreting study results, researchers should account for the site-specific characteristics leading to the initial nonrandom assignment of the treatment. Furthermore, the researcher should recognize that the altered study can no longer provide reliable knowledge of cause, but only generates hypotheses for validation when future management actions are implemented.

When replication is not possible, suspected causal effects can be masked by confounding hidden causes or by spurious correlations. Researchers may be tempted to resolve the problem by taking multiple samples as pseudoreplications. The drawback of this solution is that study results apply to study areas only and cannot be generalized to the entire system of interest.

When blocking is not feasible, precision suffers. Larger sample sizes, hence increased cost, are necessary to achieve desired levels of confidence and power. Finally, when a study considers only a portion of the system of interest (due to lack of randomization, replication, or funding), generalization of the results to the entire system could be inappropriate and misleading. In this case, researchers and managers together must re-evaluate the study objectives and scope.

Even though researchers are responsible for designing studies, managers and decision-makers should be aware of these issues and possible limitations. Other useful aspects of measurement errors are reviewed by Routledge (this volume, Chap. 5), who presents a useful set of criteria for selecting indices.

9.4 Types of Information and Study Designs

Study the past if you would divine the future.
– Confucius

Information from sources other than management experiments can play important roles in adaptive management. For example, expert judgement, anecdotes, existing data, and literature can help in building simulation models used to explore alternative scenarios and identify key uncertainties. Information from these sources can also provide supporting evidence, which becomes important when real world limitations prevent the design of "ideal" management experiments. Each source of information provides different levels of reliability.

9.4.1 Learning from existing data, expertise, and expert testimony

Using existing data and literature

In the initial stages of adaptive management, existing data and literature can be used to evaluate scenarios, project effects, or devise guidelines. However, the ability to determine treatment effects from existing data is often limited because such data may not cover the range of environments or treatments proposed, or may be knitted from disparate databases. In addition, the spatial, temporal, or ecological scope and the degree of reliability of such data may be poorly documented. Perhaps a good reminder of the potential weaknesses of using existing information is to remember the acronym for "best available data." When existing data are used, how well they can address the critical management question should be assessed honestly and accurately.

Gathering expertise and expert testimony

Another source of information is expert judgement, review, and testimony. Broad-scale assessments of wildlife population viability conducted recently by federal resource management agencies of the western United States have relied on panels of experts and contracted technical reports to fill in many gaps left by existing databases and publications (e.g., Schuster et al. 1985). In my own work using expert-panel approaches, I have modified[1] the Delphi technique (Zuboy 1981; Richey, Horner, and Mar 1985) to collect expert knowledge and judgement (Marcot et al. 1997). However, expert judgement cannot replace statistically sound experiments.

9.4.2 Learning from management actions

Probably the most reliable means of gathering information for assessing the impact of management actions is to conduct field studies. But, like publications and expert opinion, empirical evidence comes in many forms and levels of usefulness. A few key sources of evidence for the manager to know about—listed here in increasing order of reliability—include anecdotes and expert judgement, retrospective studies, nonexperimental (observational) studies, and experimental manipulation.

Anecdotes and expert judgement

The results of management actions are often evaluated informally by simple observations with no measurements. Such opportunistic observations are a two-edged foil: while the collective expertise from field experts can constitute a valuable and irreplaceable pool of wisdom, individual anecdotes can prove strikingly misleading. As a whole, anecdotal information should be used with a great deal of caution—or at least with rigorous peer review—to help avoid problems such as motivational bias (Marcot et al. 1997).

[1] Modifications addressed the need to adhere to the U.S. *Federal Advisory Committee Act,* by polling individual experts for basic ecological information and not reaching group consensus on specific management actions.

Anecdotes and expert judgement alone are not recommended for evaluating management actions because of their low reliability and unknown bias.
In the BC Forest Service, use of this source of information *alone* to evaluate management actions is not considered adaptive management.

Retrospective studies
Sometimes the results of management actions are provided by measuring the outcomes of future actions taken in the past. Retrospective studies (evaluating the outcomes of actions taken in the past) are valuable for helping to predict the outcomes of future actions. These studies can provide some insights to support or refute proposed hypotheses, and are particularly valuable for problems where some indicators take a long time to respond. However, because the treatments might not have been randomly assigned, and the initial conditions and the details of the treatments are often unknown, teasing out causal factors may be challenging at best and misleading at worst.

Nonexperimental (observational) studies
Nonexperimental studies (called observational studies by some authors) are the most common kind of field studies reported in wildlife journals. Like retrospective studies, nonexperimental studies are not based on experimental manipulations. Although it may be debatable whether nonexperimental studies should entail hypothesis testing, they should nonetheless meet statistical assumptions, including adequacy of samples sizes and selection of study sites, to ensure reliable results. Much can be learned from taking advantage of existing conditions and unplanned disturbances (Carpenter 1990; Schwarz, this volume, Chap. 3).

Nonexperimental studies usually entail analysis of correlations among environmental and organism parameters, such as studying the correlations between clearcutting and wildlife response. Causes are inferred and corroborated through repeated observations under different conditions. Because results may be confounded by uncontrolled (and unknown) factors, nonexperimental studies are best interpreted as providing only insights to cause. These insights can be valuable in predicting outcomes of actions, but again, the veracity of such predictions and the effects of management actions are best evaluated through controlled experiments (McKinlay 1975, 1985). Of nonexperimental studies, BACI-P designs allow the strongest inferences about causes (Schwarz, this volume, Chap. 3).

Inventories and *surveys* are not the same as nonexperimental studies; they display patterns but do not reveal correlates. Nevertheless, inventories and surveys can be useful in adaptive management. They provide information from which to select random samples, or a baseline of conditions from which to monitor changes over time. Inventories and surveys should still adhere to strict sampling protocols and can use more advanced statistical methods to streamline efficiencies (Schwarz, this volume, Chap. 3). For example, Max et al. (1990) presented an inventory method of random sampling of Northern Spotted Owl (*Strix occidentalis caurina*) territories with partial, annual replacement of samples to increase accuracy and reduce bias in estimates of site occupancy.

One particularly terse version of inventories is *rapid assessment procedure* (RAP) or *rapid survey*, used by some conservation groups "running ahead of the bulldozers" to survey biota of tropical forests (Oliver and Beattie 1993, 1996). Rapid surveys may prove useful in some temperate ecosystems as well, but should be used only to provide quick, initial, mostly qualitative or categorical information from which to design more formal adaptive management studies.

Experimental manipulation
Management actions can best be evaluated through experimentation (Nemec, this volume, Chap. 2). Experimental manipulations can be used to quantify the contributions from each suspected causal factor, and ultimately to develop, refine, and validate prediction models. The kind of experimentation referred to here involves deliberate, planned alterations of one or more sites, one of which may be an unaltered control.

Finally, *demonstrations* are not adaptive management per se, but often appear in the adaptive management literature (e.g., Yaffee et al. 1996). Demonstrations are designed to showcase the execution of specific management activities such as silvicultural techniques but they do not provide the evidence that controlled, replicated experiments do. When faced with a proposal for a demonstration "study," the manager might first ask if they need evidence of cause and effect, and, if so, if a management experiment with controls and replicated treatments would better provide evidence as well as the opportunity to demonstrate the activities.

9.4.3 Information for improving study designs

Study designs can be improved by using prior knowledge of the system of interest gained through retrospective analysis of past events, existing literature, and expert testimony. This information can aid in *blocking* samples to increase study efficiency, and in ensuring correct spatial and temporal representation of samples.

Study design can also benefit from *initial field sampling*. This sampling can provide preliminary estimates of variance of parameters that can be used to calculate sample size necessary to meet desired levels of precision. Initial field sampling also gives information on stratification or blocking strategy and helps to reveal conditions not originally considered in a study.

The relative merit of alternative study designs can be assessed using the tools of quantitative decision analysis, including Bayesian statistics (see Bergerud and Reed, this volume, Chap. 7; Peterman and Peters, this volume, Chap. 8). Such analysis may suggest, for example, the sampling period, sampling frequency, and sample size necessary for providing reliable information in a suitable time frame and at an acceptable cost.

The past several sections have discussed characteristics of AM study designs and use of information sources. I turn next to the topic of integrating study results into statements of risk. The topic of risk is also addressed by Peterman and Peters (this volume, Chap. 8).

9.5 Risk Analysis and Risk Management

Lots of folks confuse bad management with destiny.
– Kin Hubbard

9.5.1 Risk: speaking the same language between analysis and management

The concept of risk has pervaded much of the adaptive management literature and much of land management planning. However, researchers and managers often use the term "risk" in vastly different ways. This use can lead to, at best, confusion in interpreting results, or, at worst, misrepresentation of study results. For adaptive management, risk is defined as the expected value of adverse outcomes of a management action.

It is useful in adaptive management to differentiate risk analysis from risk management. In *risk analysis*, the researcher lists possible outcomes, estimates their likelihoods under one or more alternative future scenarios, and calculates their individual "utilities" by weighting outcome likelihoods by outcome values. These values are usually derived by managers and may pertain to social, economic, or political interests, as well as to legal regulations and objectives for resource management. Weighting outcome values with outcome likelihoods helps the manager to determine the overall risk of a management action. Then, in *risk management*, the manager defines and applies their risk attitude (their degree of risk-seeking, risk-neutral, or risk-avoidance behaviour) and then decides on the best course of action. In separating risk analysis from risk management, the onus of articulating outcome values, describing personal attitudes to risk, and defining personal decision criteria is correctly placed on the manager, not the researcher.

Formal decision analysis (Peterman and Peters, this volume, Chap. 8) is a method for assessing the risk of alternative outcomes of actions, taking uncertainty into account. Most managers do weigh the relative values or outcomes, their likelihoods, and a host of other factors that limit the decision space, such as political acceptability, effects on career, and effects on potential future decisions. However, decision analysis "in your head" is a poor substitute for quantitative decision analysis. At a minimum, managers should explicitly reveal their own outcome values, risk attidudes, and decision criteria.

9.5.2 Expressing uncertainties and unknowns

Uncertainty is a hallmark of science. However, managers—as well as politicians, the media, the public, and courts—typically view the issue of uncertainty differently than do researchers. To the researcher, uncertainty in adaptive management may represent error of measurement, confounding effects of natural variation, or other unstudied causes; such uncertainty is to be expected and results are to be treated with due care (Kodrick-Brown and Brown 1993). In some sense, the researcher may be certain of a particular level of variance, and may still view adaptive management study results as strong evidence of some effect of a management activity within some range of outcome. To the manager and others, however, such variance may be seen as lack of evidence of effects, or even as strong evidence of little or no effect, if the researcher cannot be "certain" of the outcome.

Scientific unknowns should be treated as a qualitatively different beast than scientific uncertainty. For the researcher, uncertain outcomes can be quantified as some measure of variation (such as variance or confidence interval), but unknowns cannot be quantified at all. The influence of unknowns may be deterministic, stochastic, strong, weak, or nonexistent; the researcher often simply cannot say. Again, however, the manager might erroneously view unknowns as lack of evidence of effect and thus as justification to proceed unless some contrary "proof" is provided.

Managers also need to understand how to interpret results of adaptive management studies, particularly in the context of a risk analysis. If adaptive management studies are designed as good statistical investigations, then results can serve to either falsify, or fail to falsify, the null hypothesis being tested; results can never "prove" a hypothesis.[2] Failing to falsify the null hypothesis of no effect lends only incremental credence to the management hypothesis. One of the ways to lend greater credence is through replicate findings that would further corroborate results.

Therefore, researchers and managers (as well as courts, media, and the public) must come to a common understanding of the concepts and implications of scientific uncertainty, unknowns, risk and associated concepts of proof, errors, and statistical falsification. Otherwise, results of adaptive management studies can be severely misrepresented, misunderstood, and misapplied.

9.5.3 Unravelling the causal web: when is it our fault and what can be done?

One of the main reasons for conducting adaptive management studies of resource use or ecosystem elements is to determine not just patterns and trends but also their causes. The manager should ask: What is the true cause? Do our management activities directly affect the outcome, or merely set the stage for other, more direct factors? To what degree do our management activities influence the outcome?

Untangling the causal web in field situations can be a great challenge. Seldom are causal factors affecting ecosystems single, simple, or easily quantified. Most often, factors interact in complex ways, such as with indirect and secondary effects, and through feedback relations (Figure 9.1). Even in the simplest model (Figure 9.1a), the relative contributions of known and unknown causes must be estimated. In simple models, the contribution from linear associations—which may or may not be causal—is indicated by the value of the coefficient of determination R^2 (or adjusted R^2), with the contribution from unknown associations being $1–R^2$. In more complex models, (Figures 9.1b, c, d), estimating relative contributions can be more involved. In real-world cases, it is not always evident which factors act as proximate causes, which act as less direct causes, which are mere correlates with no causal relation, and which participate in obligate feedback relations.

Of course, some relations are obvious, such as removal of forest canopy causing the local elimination of obligate canopy-dwelling organisms. But less obvious effects or gradations, though difficult to unravel, may be of great interest to the manager. For example, what degree of effect does partial removal of the forest canopy have on local plant or animal populations that are only facultatively associated with canopy environments? Might there be compounding, cumulative effects that exacerbate or ameliorate such effects, such as wider regional loss of forest canopies, or restoration of canopy conditions in adjacent stands?

To determine the relative influence of specific management activities, the researcher may turn to statistical techniques using estimation of partial correlations. These methods help determine the contribution of one factor, such as a management activity, given the effects of all other factors (e.g., other activities, natural changes in environments, unknown causes). Traditional analyses such as step-wise multiple regression help identify such partial influences. Other, less well-known techniques such as regression trees and path regression analysis (e.g., Schemske and Horvitz 1988) can also be used.

Determining the relative influence of management actions is vital for setting realistic expectations for management results. For example, determining that fragmentation of local forests affects breeding habitat for migrating songbirds (Wilcove 1985) is only part of the puzzle; loss of habitat on neotropical wintering grounds is also a significant cause of declines in songbird populations. Therefore, changing local management to reduce fragmentation should be expected to have only a partial impact on songbird populations.

[2] Some authors suggest that Bayesian analyses also can be interpreted as the testing of null hypotheses, that is, the prior probabilities.

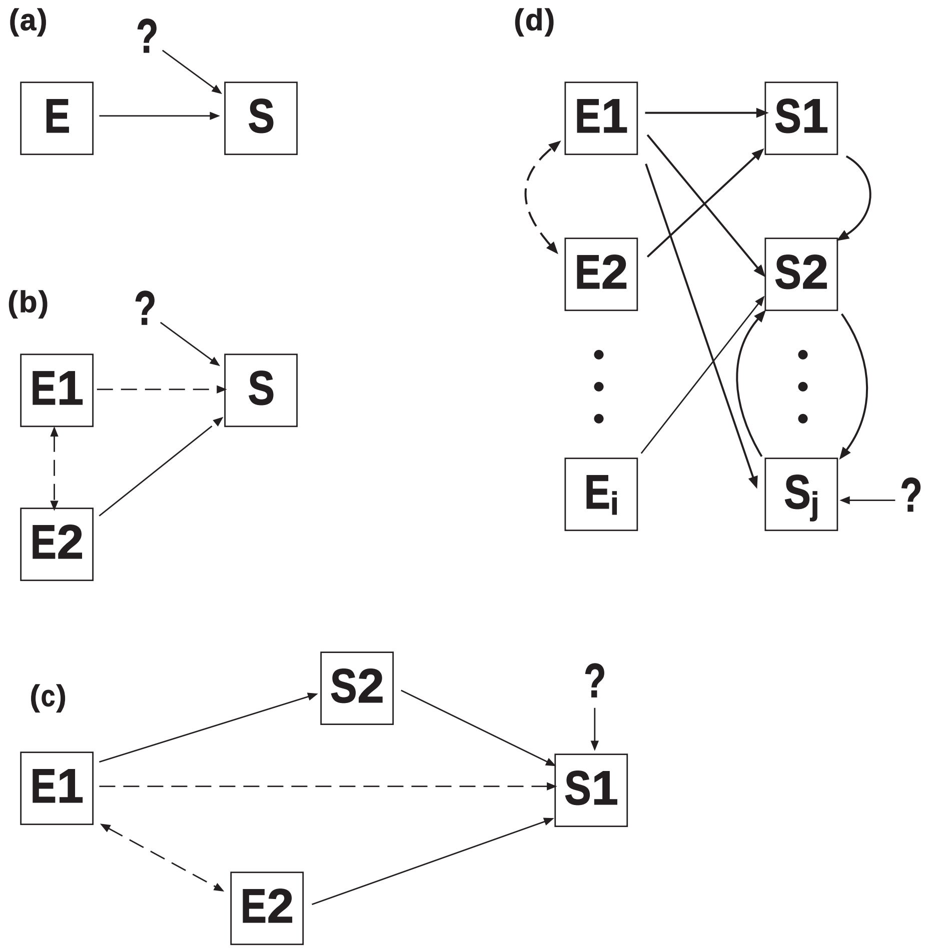


FIGURE 9.1 *Causes and correlates: four examples. In all figures, S = wildlife species response; ? = unexplained variation due to measurement error, experimental error, or effects of other environmental or species factors; solid arrows = causal relations; dotted arrows = correlational relations that may or may not be causal. (a) In this simplest case, some wildlife species response S, such as population presence or abundance, is assumed to be explained and caused by some environmental factor E. (b) In a more complex case, we may be measuring one environmental factor E1 when the real cause is another environmental factor E2. (c) Getting closer to the real world, a second species response S2 may be part of the cause. (d) Most like the real world, with feedback relations among the dependent (response) variables S. (Adapted from Morrison et al. 1998, Fig. 10.2.)*

9.5.4 A dilemma for managers: when samples are few and crises are many

One bane of adaptive management is that, in many cases, the unique field conditions make it difficult to correctly design statistical studies to identify and quantify causes. Especially when studying landscapes, ecosystems, rare or threatened species, and infrequent events, the major problems in the design of such studies are small sample size and inability to replicate conditions. In such circumstances, what can the researcher do, and how should the manager interpret results? The answer may be found in several courses of action: selecting correct indicators, merging disparate lines of evidence, and using statistical procedures that take advantage of prior knowledge or that function adequately with small sample sizes.

Selecting correct indicators

Indicators that are objective, repeatable measurements, whose quality is documented quantitatively should be selected. For adaptive management studies, an indicator should (1) respond rapidly to changes, (2) signal changes in other variables of interest, (3) be monitored efficiently, and (4) be causally linked to changes in stressors. Most "ecological indicators" purported to fit these criteria usually fail (Block et al. 1987; Patton 1987; Landres et al. 1988). For example, the Northern Spotted Owl, often selected by USDA Forest Service as an "old-growth indicator," may serve criterion (4), but fails with the other three criteria: spotted owls have low reproductive rates and long life spans, so they respond slowly to changes; changes in their populations may not necessarily correlate well with other desired facets of old-growth forests (e.g., habitat for anadromous fish); and their population trends are terribly costly to monitor.

Indicators that do meet these criteria include soil arthropods as indicators of soil productivity (McIver et al. 1990; Moldenke and Lattin 1990; Pankhurst et al. [editors] 1997); butterfly diversity as an indicator of overall ecosystem diversity (Kremen 1994); and some arboreal lichens as indicators of air quality (Stolte et al. 1993; Geiser et al. 1994) or persistence of old forests (Tibell 1992). See Murtaugh (1996) for a review of the statistical basis of ecological indicators.

Merging disparate lines of evidence

Merging different study results is a second tactic that can help in identifying causal relations when good experimental design is impossible or impractical. In statistics, this process is called "combining information" (CI). Draper et al. (1992) provide a useful overview of various CI techniques, including methods of meta-analysis (Hedges and Olkin 1985) that can be useful in conservation research (Fernandez-Duque and Valeggia 1994). For example, meta-analysis was used by Burnham et al. (1996) to determine overall trends of Northern Spotted Owls by combining results from individual population demography studies.

CI is not a panacea, as it can be fraught with difficulties such as matching consistency and representativeness among studies designed for different initial objectives. Still, the researcher may wish to use CI methods to merge lines of evidence taken from available information. This available information could include anecdotes and local experience, retrospective studies, observational studies, experimental manipulations, and demonstrations. The reliability of each source for inferring causes should be judged very carefully.

In contrast to formal meta-analysis, simply pooling data from different studies could lead to spurious and misleading conclusions. For example, to assess the impact of clearcutting on grizzly bear populations, the data from several studies might be combined into an overall regression. This regression might suggest a significant correlation between clearcutting and grizzly bear populations. However, grizzly bears within individual study areas might respond differently to clearcutting because they come from different geographic areas, latitudes, or forest types. Thus the correlation may reflect these differences between populations, rather than any treatment effect. The incorrect conclusion of correlation would arise because such an analysis violates an assumption underlying regression: that the data come from the same statistical population with the same causal mechanisms. On the other hand, a formal meta-analysis approach would analyze results from each study with differences among studies as an explanatory factor. CI has great utility, especially where powerful experimental studies are difficult. However, managers and researchers must be careful in its use, ensuring that studies are truly from the same causal web.

Using statistical procedures that take advantage of prior knowledge

Bayesian statistics were developed specifically for using prior knowledge and incrementally gathered field data (Ver Hoef 1996). Bayesian statistical techniques include empirical Bayes and sequential Bayes procedures, in which initial estimates of the likelihood of conditions become incrementally adjusted and refined over time as new evidence is gathered (e.g., Gazey and Staley 1986; Link and Hahn 1996). Expert opinion, existing literature and data, retrospective studies, and non-experimental studies can all be used to establish preliminary values of prior probabilities in a Bayesian analysis. Bayesian methods were reviewed by Bergerud and Reed (this volume, Chap. 7), who advocate their use to incorporate accumulated knowledge of experts.

9.6 Conclusions and Recommendations

Knowledge is the small part of ignorance that we arrange and classify.

– Ambrose Bierce

9.6.1 A decision tree for managers

The six stages of adaptive management and sources of information appropriate for each stage are presented in Table 9.1. This table can be used by managers as a decision tree to guide the choice of

TABLE 9.1 *Stages of an adaptive management (AM) project and sources of information appropriate for each stage. This table can be used by managers as a decision tree to guide (1) the choice of study for each AM stage (reading across rows), and (2) the use of existing information (reading down columns). ✓ = recommended sources; ✗ = not recommended; [✓] = most recommended for a given project stage; – = does not apply; (pilot) = use for pilot study only. See Chapter 1 for full descriptions of AM stages.*

AM stages	Literature review	Expert judgement	Demonstration	Anecdote	Retrospective study	Nonexperimental study	Experimental study
1. Assess problem Identify potential impacts of management actions and the potential reasons for them.							
Identify patterns and trends	✓	✓	✗	✓	✓	✗	✗
Identify correlates	✓	✓	✗	✓	✓	✗	✗
Identify potential causes of suspected impact [a]	✓	✓	✗	✓	✓	✗	✗
2. Design project Determine treatments to implement, sample size, effect size, power, etc.	✓	✓	✗	✗	✓	[✓][b] (pilot)	[✓][b] (pilot)
3. Implement	—	—	—	—	—	✓	[✓]
4. Monitor	Monitoring design is determined at the “Design project” stage.						
5. Evaluate Interpretation	✓[c]	✓[c]	✗	✗	✓[c]	✓	[✓]
6. Adjust management action	—	—	—	—	—	—	—

[a] Experimental and nonexperimental studies can provide information on patterns, correlates, etc., but typically these studies will not be done by taking advantage of management actions, but rather as part of applied research.

[b] All else being equal, if the cost of conducting a nonexperimental study is significantly less than that of an experimental study, choose the former.

[c] In the “Evaluation” stage, existing information based on literature, expert judgement, and retrospective analysis is updated using data collected from the management experiment to assess the effect or outcome of an action. It can also be used to determine the relative plausibility of suspected causes, and to estimate the prior probabilities in a Bayesian analysis.

study for each stage of adaptive management, as well as to guide the use of existing information.

At the problem assessment stage, existing information is valuable for identifying *potential* impacts of management actions. At the project design stage, pilot studies (experimental or nonexperimental) are recommended for fine tuning the study methodology. Pilots can be used to estimate variability in the response variables; these estimates can then be used to determine sample size, effect size, and power for the study. Controlled experiments allow the strongest inference about the actual impacts of management actions. Once a study has been implemented, relevant data are collected through a monitoring process. The data are then analyzed using appropriate statistical methods to answer questions set out at the beginning of the adaptive management project. In the evaluation stage, existing knowledge (based on literature, expert judgement, and retrospective analysis) is updated using data collected from the management experiment to assess the effect or outcome of an action. Using Bayesian analysis, existing knowledge together with collected data can also be used to determine the relative plausibility of suspected causes. Management actions are then adjusted based on this updated knowledge. During the course of the management experiment, new questions may arise that then lead to further problem assessment, project design, implementation, and so on, in a continuous cycle of learning.

9.6.2 Is there a "best" statistical approach to adaptive management?

The answer to this question is an unqualified "yes." The best approach for answering the questions "Did this action have the desired effect?" and "Are the basic assumptions underlying management decisions correct?" is to use controlled, randomized experiments with sufficient sample sizes and duration. This approach provides the best understanding of causal relations and the best basis for validating predictive models—assuming that the models can provide testable predictions.[3] Nevertheless, the informativeness of a statistical approach must also be weighed against its costs (ecological, social, and economic). Ultimately, designing management actions as controlled, randomized experiments will provide the best evidence for managers who face the difficult task of making management decisions and defending such decisions legally, politically, and scientifically.

Short of this ideal, both researchers and managers have their work cut out for them. They should maximize the use of available information, but not draw undue conclusions about causes. It may be useful to explicitly array the various available lines of evidence and to articulate the confidence in identifying causes from each. Managers and researchers must look for similarities and disparities among lines of evidence and investigate reasons for the differences. Moreover, they should seek peer review to ensure appropriate and rigorous use of various sources of information. Repeatability of findings and suspected causes is the basis for true scientific understanding and predictability.

Real-world adaptive management problems are often complicated by time exigencies or finite funding so that powerful experiments are not possible. If the study design must be compromised, then the ramifications of drawing incorrect conclusions should be thought out. When turning to field experts for their professional opinions, managers should be aware of potential problems such as motivational and personal bias.

Managers should weigh the benefits and cost of a more or less rigorous approach. For, in the end, reliable knowledge is a hard-won but essential commodity for ensuring successful conservation practices for future generations.

Acknowledgements

My thanks to Brian Nyberg and Brenda Taylor for suggesting topics to cover in this chapter and for their technical reviews of the manuscript. Roger Green, Rick Page, Martin Raphael, and Ian Thompson also provided technical reviews. Thanks to Vera Sit and Brenda Taylor for their fine edits. My gratitude to Tim Max for his technical review and for thoughtful discussions on the role of statistics in adaptive management.

[3] As expressed by one statistician, if predictive models are so complex that they become essentially untestable, then they are nothing more than belief structures and their relation to science is questionable at best (T. Max, pers. comm., 1997).

References

Anderson, J.L. [n.d.]. Errors of inference. This volume.

Block, W.M., L.A. Brennan, and R.J. Gutierrez. 1987. Evaluation of guild-indicator species for use in resource management. Environ. Manage. 11:265–9.

British Columbia Ministry of Forests. 1996. Adaptive management: learning from our forests. Brochure. Integrated Resources Policy Branch, Victoria, B.C.

Burnham, K.P., D.R. Anderson, and G.C. White. 1996. Meta-analysis of vital rates of the northern spotted owl. Stud. Avian Biol. 17:92–101.

Carpenter, S.R. 1990. Large-scale perturbations: opportunities for innovation. Ecology 71:2038–43.

Collopy, M., J. Cannell, S. Daniels, D. Dippon, E. Gaar, G. Grant, B. Mulder, C. Philpot, J. Steffenson, and F. Swanson. 1993. Implementation and adaptive management. *In* Forest ecosystem management: an ecological, economic, and social assessment. Forest Ecosystem Management Assessment Team. U.S. Dep. Agric. For. Serv., Washington, D.C.

Draper, D., D.P. Gaver, Jr., P.K. Goel, J.B. Greenhouse, L.V. Hedges, C.N. Morris, J.R. Tucker, and C.M. Waternaux. 1992. Combining information: statistical issues and opportunities for research. National Academic Press, Washington, D.C. Contemporary statistics No. 1.

Fernandez-Duque, E. and C. Valeggia. 1994. Meta-analysis: a valuable tool in conservation research. Cons. Biol. 8:555–61.

Gazey, W.J. and M.J. Staley. 1986. Population estimation from mark-recapture experiments using a sequential Bayes algorithm. Ecology 67:941–51.

Geiser, L.H., C.C. Derr, and K.L. Dillman. 1994. Air quality monitoring on the Tongass National Forest: methods and baselines using lichens. U.S. Dep. Agric. For. Serv., Alaska Region. Petersburg, Alaska. R10-TB-46.

Gentiol, S. and G. Blake. 1981. Validation of complex ecosystem models. Ecol. Modelling 14:21–38.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. J. Wiley, New York, N.Y.

Hedges, L.V. and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, New York, N.Y.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211.

Kodric-Brown, A. and J. H. Brown. 1993. Incomplete data sets in community ecology and biogeography: a cautionary tale. Ecol. Applic. 43:736–42.

Kremen, C. 1994. Biological inventory using target taxa: a case study of the butterflies of Madagascar. Ecol. Applic. 4:407–22.

Landres, P.B., J. Verner, and J.W. Thomas. 1988. Ecological uses of vertebrate indicator species: a critique. Cons. Biol. 2:316–28.

Ligon, J.D. and P.B. Stacey. 1996. Land use, lag times and the detection of demographic change: the case of the acorn woodpecker. Cons. Biol. 10:840–6.

Link, W.A. and D.C. Hahn. 1996. Empirical Bayes estimation of proportions with application to cowbird parasitism rates. Ecology 77:2528–37.

Loehle, C. 1990. Indirect effects: a critique and alternate methods. Ecology 71:2382–6.

McIver, J.D., A.R. Moldenke, and G.L. Parsons. 1990. Litter spiders as bio-indicators of recovery after clearcutting in a western coniferous forest. NW Environ. J. 6:410–2.

McKinlay, S.M. 1975. Comprehensive review on design and analysis of observational studies. J. Am. Statist. Assoc. 70:503–20.

______. 1985. Observational studies. *In* Encyclopedia of statistical sciences. Vol. 6. J. Wiley, New York, N.Y. pp. 397–401.

Marcot, B.G. 1986. Summary: biometric approaches to modeling: the manager's viewpoint. *In* Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. J. Verner, M.L. Morrison, and C.J. Ralph (editors). Univ. Wis. Press, Madison, Wis. pp. 203–4.

Marcot, B.G., M.A. Castellano, J.A. Christy, L.K. Croft, J.F. Lehmkuhl, R.H. Naney, R.E. Rosentreter, R. E. Sandquist, and E. Zieroth. 1997. Terrestrial ecology assessment. *In* An assessment of ecosystem components in the interior Columbia Basin and portions of the Klamath and Great Basins. T.M. Quigley and S.J. Arbelbide (editors). Pac. N. Res. Sta., Portland, Oreg. U.S. Dep. Agric., For. Serv. Gen. Tech. Rep. PNW-GTR-405. pp. 1497–1713.

Max, T.A., R.A. Souter, and K.A. O'Halloran. 1990. Statistical estimators for monitoring spotted owls in Oregon and Washington in 1987. U.S. Dep. Agric. For. Serv., Portland Oreg. Res. Pap. PNW-RP-420.

Moldenke, A.R. and J.D. Lattin. 1990. Density and diversity of soil arthropods as "biological probes" of complex soil phenomena. NW Environ. J. 6:409–10.

Morrison, M.L., B.G. Marcot, and R.W. Mannan [1998]. Wildlife-habitat relationships: concepts and applications. 2nd ed. Univ. Wis. Press, Madison, Wis. In press.

Murtaugh, P.A. 1996. The statistical evaluation of ecological indicators. Ecol. Applic. 6:132–9.

Nemec, A.F.L. [n.d.]. Design of experiments. This volume.

Oliver, I. and A.J. Beattie. 1993. A possible method for the rapid assessment of biodiversity. Cons. Biol. 7:562–8.

______. 1996. Designing a cost-effective invertebrate survey: a test of methods for rapid assessment of biodiversity. Ecol. Applic. 6:594–607.

Pankhurst, C., B. Doube, and V. Gupta (editors). 1997. Biological indicators of soil health. Oxford Univ. Press, New York, N.Y.

Patton, D.R. 1987. Is the use of "management indicator species" feasible? W. J. For. 2:33–4.

Peterman, R.M. and C. Peters. [n.d.]. Decision analysis: taking uncertainties into account in forest resource management. This volume.

Richey, J.S., R.R. Horner, and B.W. Mar. 1985. The Delphi technique in environmental assessment. II. Consensus on critical issues in environmental monitoring program design. J. Environ. Manage. 21:147–59.

Richey, J.S., B.W. Mar, and R.R. Horner. 1985. The Delphi technique in environmental assessment. I. Implementation and effectiveness. J. Environ. Manage. 21:135–46.

Routledge, R.D. [n.d.]. Measurements and estimates. This volume.

Schemske, D.W. and C.C. Horvitz. 1988. Plant-animal interactions and fruit production in a neotropical herb: a path analysis. Ecology 69:1128–38.

Schuster, E.G., S.S. Frissell, E.E. Baker, and R.S. Loveless. 1985. The Delphi method: application to elk habitat quality. U.S. Dep. Agric. For. Serv. Res. Pap. INT-353.

Schwarz, C.J. [n.d.]. Studies of uncontrolled events. This volume.

Smith, G.J. [n.d.]. Retrospective studies. This volume.

Stolte, K., D. Mangis, R. Doty, and K. Tonnessen. 1993. Lichens as bioindicators of air quality. Rocky Mtn. For. Range Exp. Sta., Fort Collins Colo. U.S. Dep. Agric. For. Serv. Gen. Tech. Rep. RM-224.

Thomas, L. and K. Martin. 1996. The importance of analysis method for breeding bird survey population trend estimates. Cons. Biol. 10:479–90.

Tibell, L. 1992. Crustose lichens as indicators of forest continuity in boreal coniferous forests. Nor. J. Bot. 12:427–50.

Tiebout, H.M. III and K.E. Brugger. 1995. Ecological risk assessment of pesticides for terrestrial vertebrates: evaluation and application of the U.S. Environmental Protection Agency's quotient model. Cons. Biol. 9:1605–18.

Ver Hoef, J.M. 1996. Parametric empirical Bayes methods for ecological applications. Ecol. Applic. 6:1047–55.

Wilcove, D.S. 1985. Forest fragmentation and the decline of migratory songbirds. Ph.D. thesis, Princeton Univ., Princeton, N.J.

Yaffee, S.L., A.F. Phillips, I.C. Frentz, P.W. Hardy, S.M. Maleki, and B.E. Thorpe. 1996. Ecosystem management in the United States. Island Press, Covelo, Calif.

Zuboy, J.R. 1981. A new tool for fishery managers: the Delphi technique. N. Am. J. Fish. Manage. 1:55–9.

A posteriori**:** Referred to after the data have been collected and examined.

A priori**:** Referred to before the data are collected and examined.

Accuracy: The nearness of a measurement to the actual value of the variable being measured.

Active adaptive management: Management is designed as an experiment to compare alternative actions (treatments) or discriminate among alternative hypotheses about how the system responds to actions. Active adaptive management can involve deliberate "probing" of the system to identify thresholds in response and clarify the shape of the functional relationship between actions and response variables.

Alternative hypothesis: A claim or research hypothesis that is compared with another (usually null) hypothesis.

Analysis of variance (ANOVA): A group of statistical procedures for analyzing continuous data sampled from two or more populations, or from experiments in which two or more treatments are used. ANOVA procedures partition the variation observable in a response variable into two basic components: (1) variation due to assignable causes and (2) uncontrolled or random variation. Assignable causes refer to known or suspected sources of variation from variates that are controlled (experimental factors) or measured (covariates) during an experiment. Random variation includes the effects of all other sources not controlled or measured during the experiment.

Analytical survey: A type of nonexperimental study where groups sampled from a population of units are compared.

Autocorrelation: Occurrence when consecutive measurements in a series are not independent of one another. Also called serial correlation.

Bayes decision: The optimal decision identified when uncertainties are considered using a formal decision analysis.

Bias: The deviation of a statistical estimate from the quantity it estimates. Bias can be a systematic error introduced into sampling or testing. Positive bias will overestimate the parameter; negative bias will underestimate it.

Blocking: A design technique where experimental units are grouped into homogeneous blocks, according to some identifiable characteristic(s). Successful blocking reduces the experimental error that results from variation among heterogeneous units.

Conditional probability, P(A|B): The probability of the event A given that a related event B has taken place.

Confidence limits: Confidence limits indicate the precision of a parameter estimate. If samples of size *n* were repeatedly obtained from the population and constructed $(1-\alpha)\%$ confidence limits for each, the expected result would be that $100(1-\alpha)$ out of 100 confidence limits would contain the true parameter.

Confounding: Confounding occurs when one or more effects cannot be unambiguously attributed to a single factor or interaction.

Control: A treatment level included in an experiment to show what would have happened if no treatments had been applied to the experimental material.

Controlled experiment: An experiment in which the experimenter controls the treatments to be compared, and can randomly assign experimental units to the treatments. Also called a designed experiment.

Correlation: A measure of the strength of the linear relationship between two random variables. A strong correlation between two random variables does not necessary signify a causal relationship between them.

Covariate: A variable that influences the response but is unaffected by any other experimental factors. Including a covariate in the analysis may increase the power of the analysis to detect treatment effects.

Decision analysis: A structured, formalized method for ranking management actions that are being considered. It quantitatively takes into account uncertainties.

Delphi technique: A procedure for interviewing experts and capturing their expertise by striving to reach consensus in an expert panel or group setting.

Effect size: The treatment effect the experimenter wants to be able to detect. Effect size influences the statistical power of an analysis: larger effect size yields greater statistical power.

Expected value of an outcome: The weighted average outcome, where each outcome is weighted by the probability assigned to its branch on the decision tree. Summing these weighted outcomes for each management action gives the "expected value" of that action.

Experimental design: A plan for assigning treatments to experimental units and the statistical analysis associated with the plan. It includes formulation of statistical hypotheses, choice of experimental conditions, specification of the number of experimental units required and the population from which they are to be sampled, assignment of treatments to experimental units, determination of the dependent variables to be measured, and the statistical analysis to be performed.

Experimental error: Any variation, including sampling (measurement) error and natural variation error, that cannot be explained by the experimental factors.

Experimental factor: Any treatment or variable that is controlled in an experiment, either by physically applying a treatment to an experimental unit or by deliberately selecting a unit with a particular characteristic.

Experimental unit: The entity to which one treatment (level of one or more factors) is applied. Also called a treatment unit.

Explanatory variable: A variable that is thought to provide information on the value of the response variable.

Homogeneous: Experimental units are homogeneous when they do not differ from one another in any systematic fashion and are as alike as possible on all characteristics that might affect the response.

Hypothesis: A tentative assumption, adopted to account for certain facts that can be tested.

Hypothesis testing: A type of statistical inference for assessing the validity of a hypothesis by determining whether it is consistent with the sample data.

Impact survey: A type of nonexperimental study where one site affected by some planned or unplanned event is compared with a control site not affected by the event. Impact surveys are typically used to investigate the effects of large-scale, unreplicated events. Types of impact surveys include BACI (Before-After-Control-Impact) where variables in both an impact (treatment) and control site are compared before and after some event, and BACI-P (Before-After-Control-Impact-paired)—an extension of BACI where control and impact sites are sampled at the same points in time, both before and after the event.

Null hypothesis: A statistical hypothesis that states that there is "no difference" between the true value of a parameter and the hypothesized value, or that there is "no effect" of a treatment.

Observational survey: A type of nonexperimental study where results from two units or sites are compared. Results are applicable only to the units sampled, and cannot be extrapolated to other units or sites.

Parameter: A numerical characteristic of a population. It is often estimated by a sample statistic.

Passive adaptive management: Managers implement what they assume, based on existing information, is the "best" action (i.e., the action most likely to produce the desired outcome). Adjustments are made when actual outcomes deviate from predictions. The limitation of passive adaptive management is that it can be difficult to determine why actual outcomes deviate from predictions.

Power (1–β): The probability of correctly rejecting a null hypothesis when it is actually false. Also called statistical power, or the power of a test.

Precision: The closeness to each other of repeated measurements of the same quantity. Precision should not be confused with accuracy. Imagine a dart board: accuracy refers to the distance of the dart from the bull's-eye; precision refers to how tightly grouped repeated dart throws are.

Prospective study: A study where actions (treatments) have not yet been applied, and data have not yet been collected. Prospective studies may be either experimental or nonexperimental. Contrast with retrospective study.

Pseudoreplication: Refers to various violations of the assumption that replicated treatments are independent. A common form of pseudoreplication occurs when multiple subsamples from one treatment unit, rather than samples from multiple (replicated) treatment units, are used to calculate the statistical probability of a treatment effect.

P-value: Probability of obtaining a value for a test statistic that is as extreme as or more extreme than the observed value, assuming the null hypothesis is true. In classical hypothesis testing, the null hypothesis is rejected when the P-value is less than the chosen significance level (α).

R^2: A statistic that assesses how well a regression model describes the relationship between the dependent and independent variables. For comparisons of several models using the same data, it is more appropriate to use the adjusted R^2—R^2 adjusted by the model degrees of freedom.

Randomization: Treatments are randomly assigned to the experimental units, so that each unit has a known and independent chance of being allocated a particular treatment. Randomization protects against possible bias (systematic error) by ensuring that all unmeasured factors are more or less evenly distributed among treatments.

Random sampling: A scheme for choosing subjects from a population, so that each member of the population has a known (often equal) and independent chance of being selected. Random sampling allows you to generalize the results of the experiment to the population from which the sample was drawn.

Regression: A relationship where the magnitude of one variable (the dependent variable) is determined in part by the magnitude of another variable (the independent variable).

Replication: Replication involves applying the same combination of factors to more than one experimental unit. Replication is a means of assessing the variability that is not attributable to the treatment.

Response variable: A variable measured to assess the outcome of an experiment. In regression, the response variable is referred to as the dependent variable.

Retrospective study: A study that uses data already collected for other purposes or actions (treatments) that have already been implemented. A retrospective study is a type of uncontrolled (non-experimental) study. Contrast with prospective study.

Risk: The expected loss associated with an outcome or decision. Risk is the product of the possible magnitude of a loss and the probability of it occurring.

Sample: A subset of measurements or observations taken from a population. Conclusions about the characteristics of the population can be drawn from the characteristics of the sample.

Sampling design: A plan that describes the nature of the sampling units, the number of sampling units, the method of selection, and the variables to be measured.

Sampling unit: A basic unit selected for sampling.

Scientific (research) hypothesis: A testable proposition that is tentatively adopted to account for observed facts and to guide investigation.

Sensitivity analysis: A procedure for assessing the degree to which predicted outcomes vary with changes in assumptions about parameter values.

Significance level (α): The probability of making a Type I error (i.e., rejecting a true null hypothesis). In hypothesis testing, indicates the maximum amount of Type I error the experimenter is willing to tolerate.

Standard deviation: A measure of the dispersion (variability) of the data. The deviations of individual observations from the sample mean are squared, the squares are averaged, and the square root of the result is calculated.

Standard error: The standard deviation of a sample statistic. Standard deviation is a measure of the dispersion of the individual observations from their mean; standard error is a measure of the dispersion of repeated sample statistics from their mean.

Statistic: A numerical characteristic that is computed from a sample of observations and that estimates a population parameter.

Statistical hypothesis: It states the scientific hypothesis in precise, quantitative terms, often as a variable whose sampling distribution can be described statistically.

Statistical independence: Observations are statistically independent if the value of one of the observations does not influence the value of any other observations. Simple random sampling produces independent observations.

Statistical inference: The act of drawing a conclusion about the characteristics of a population by analyzing the characteristics of a sample (i.e., generalizing about the whole, based on information from a part of the whole).

Stochastic process: A process that is not completely predictable.

Stratification: Survey units are grouped into homogeneous groups, according to some identifiable characteristic(s). Each stratum is then surveyed. Stratification in sampling is analogous to blocking in experimental design.

Systematic sampling: A sampling scheme where every k^{th} unit in a population is sampled (with the result that sampling points are a fixed distance apart).

Type I error: The error of rejecting a null hypothesis that is true.

Type II error: The error of not rejecting a null hypothesis that is false.

Uncontrolled experiment: An experiment in which the investigator does not control the selection of treatments or the assignment of treatments to the experimental units.

Variance: A measure of the dispersion (variability) of the data. The variance is the square of the standard deviation.