CONTEMPORARY BAYESIAN AND FREQUENTIST STATISTICAL RESEARCH METHODS FOR NATURAL RESOURCE SCIENTISTS

Howard B. Stauffer
Mathematics Department, Humboldt State University, Arcata, California

CONTEMPORARY BAYESIAN AND FREQUENTIST STATISTICAL RESEARCH METHODS FOR NATURAL RESOURCE SCIENTISTS

The Wiley Bicentennial–Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce
President and Chief Executive Officer

Peter Booth Wiley
Chairman of the Board

CONTEMPORARY BAYESIAN AND FREQUENTIST STATISTICAL RESEARCH METHODS FOR NATURAL RESOURCE SCIENTISTS

Howard B. Stauffer
Mathematics Department, Humboldt State University, Arcata, California

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:

Stauffer, Howard B., 1941-
Contemporary Bayesian and frequentist statistical research methods for natural resource scientists/Howard B. Stauffer.
p. cm.
ISBN 978-0-470-16504-1 (cloth)
1. Bayesian statistical decision theory. 2. Mathematical statistics. I. Title.
QA279.5.S76 2008
519.5'42—dc22

2007015575

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my parents,
Howard Hamilton Stauffer and Elizabeth Boyer Stauffer,
and to my family,
wife Rebecca Ann Stauffer,
daughter Sarah Elizabeth Stauffer,
and son Noah Hamilton Stauffer.
Their love and support has sustained me and provided
meaning and joy in my life.

CONTENTS

PREFACE

This book began as a critique against the current misuses of statistics in the natural resource sciences. I had worked for many years as a forestry and wildlife management statistician, in academia, government, and industry. I was frustrated with the frequent misuse of statistical analysis and inference with natural resource data. Hypothesis testing was commonly misused with observational data to compare so-called habitat treatments such as old-growth and young-growth forest habitat for their effects on wildlife species. Such hypotheses were statistical rather than scientific, referring to specific stands of interest. Many null hypotheses were "silly" and clearly not true. Sample datasets were not completely randomized, and "experimental" conditions were not effectively controlled. I was reviewing manuscripts and attending seminars where null hypotheses were being rejected that were clearly false a priori, and effect sizes between treatments, the differences of biological importance that were of interest to wildlife managers, were not even being estimated. More seriously, null hypotheses that were clearly false were being "supported" by hypothesis testing results that failed to reject, in studies where sample sizes were small, effect sizes of importance, and power to detect these effect sizes were not specified, and this power was very likely small.

Natural resource scientists did not clearly understand how to interpret their inferences from frequentist statistical analysis. The indirect logic of frequentist statistical inference, in interpreting the meaning of confidence intervals or the test statistics and p values from hypothesis testing, was proving to be very confusing to natural resource scientists. The challenge of natural resource scientists in explaining such frequentist inferences to managers, attorneys, politicians, and the public was proving to be even more daunting.

I was concerned with the extent of data dredging that was common in the field. I was commonly seeing datasets collected for habitat selection modeling using multiple linear regression or logistic regression analysis with measurements for over 100 covariates and sample sizes under 100. Scientists were not giving enough thought to sampling design and the type of analysis appropriate for their studies, prior to data collection. Stepwise and best-subsets selection methods were being utilized without concern for their potential for overfitting sample datasets with large and unspecified amounts of compounded error.

Then, around 1998/99, several pioneering applied statisticians with many years of experience in the field of wildlife management began to show the way out of this wilderness. Ken Burnham and David Anderson (1998, 2002) published their

landmark book advocating the use of a priori model selection and inference using the Akaike information criterion (AIC) as a way of reducing model overfitting and compounding of error with model selection and inference. Doug Johnson's (1999) article critiquing the misuses of hypothesis testing in wildlife management research was published in the Journal of Wildlife Management. These ideas took the wildlife management research community by storm. Ray Hilborn and Mark Mangel (1997) published a seminal book advocating the use of Bayesian statistical analysis and inference in the fields of ecology and fisheries management. Pinhiero and Bates (2000) published an important book describing the applications of mixed-effects modeling in S-Plus. Ramsey and Schafer (2002) warned natural resource scientists about the distinctions between observational and experimental data. Because of some of these influences, a priori model selection and inference has become the accepted dominant paradigm for model selection and inference in the field of wildlife management. The misapplication of hypothesis testing has been reduced. Perhaps even too hastily, the old ways of doing statistics have been discarded in the rush to remain "current." Meanwhile, many other important contemporary methods of applied statistics remain relatively unknown among natural resource scientists, methods such as generalized linear modeling, mixed-effects modeling, and Bayesian statistical analysis and inference.

This book was written to introduce these newer contemporary methods of statistical analysis to natural resource scientists and strike a balance between the old and new ways of doing statistics. Chapter 1 introduces three case studies that illustrate the need for newer contemporary methods of statistical analysis and inference for natural resource science applications. It also reviews some of the most important fundamental methods of traditional frequentist statistical analysis and inference and ends with a brief introduction to the frequentist software S-Plus and R that are used throughout the book. Chapters 2–4 introduce an alternative approach to traditional frequentist statistical analysis and inference, namely, Bayesian statistical analysis and inference. These three chapters provide an introduction to the fundamental concepts of Bayesian statistical analysis, its historical background, conjugate solutions, Bayesian hypothesis testing and decisionmaking, Markov Chain Monte Carlo (MCMC) solutions, and applications in WinBUGS (Windows version of Bayesian statistical inference Using Gibbs Sampling) software. Chapter 5 presents two alternative strategies to model selection and inference, a posteriori model selection and inference, and a priori parsimonious model selection and inference using AIC and the deviance information criterion (DIC). Chapter 6 introduces the ideas of generalized linear modeling (GLM), focusing on the most popular GLM of logistic regression. Chapter 7 presents an introduction to mixed-effects modeling in S-Plus® and R. Chapters 5–7 provide applications with both frequentist and Bayesian statistical analysis and inference approaches, illustrating the strengths and limitations of each approach. Chapter 8 concludes with a summary of the contemporary methods introduced in this book.

This book can be used as a textbook for an intermediate undergraduate or introductory graduate semester course in contemporary research statistics for natural resources sciences. It assumes a minimum prerequisite undergraduate course in introductory statistics that includes the estimation of parameters such as mean and proportion;

hypothesis testing with t tests, F tests for analysis of variance (ANOVA), and chi-square (χ^2) tests; and linear regression analysis. Parts of the book can be read independently along with the introductory Chapter 1, Chapters 2–4 on Bayesian statistical analysis and inference, Chapter 5 on strategies for model selection and inference, Chapter 6 as an introduction to generalized linear modeling, and Chapter 7 on mixed-effects modeling. The book can also be read and used as a refresher manual or a reference book by natural resource scientists. Parts of the book have served as resource materials that I have used for 2-day workshops on topics of statistics such as Bayesian statistical analysis and inference using WinBUGS and capture–recapture analysis using MARK.

I'd like to thank many colleagues who have provided advice, encouragement, and support throughout my career as an applied statistician and influenced a perspective that has led to the writing of this book: David Anderson, Doug Johnson, Barry Noone, Bill Zielinski, C. J. Ralph, Cindy Zabel, Hart Welsh, Cynthia Perrine, Larry Fox, Jan Derksen, Rich Padula, Bryan Gaynor, Ken Mitchell, Sam Otukol, A. Y. Omule, David Gilbert, Les Safranyik, Mark Rizzardi, Yoon Kim, Butch Weckerly, David Hankin, John Sawyer, Mike Messler, Andrea Pickart, Matt Johnson, and Mark Colwell. I'd also like to thank the Wiley staff who were so helpful during the publication process: editor Susanne Steitz-Filler, senior production editor Kris Parrish, and copy editor Cathy Hertz. My career has been a most interesting one, working with natural resource scientists in academia, government, and industry, applying traditional and contemporary ideas in the application of statistical design and analysis to the natural resource sciences. It is up to natural resource scientists to make the most appropriate and effective choices on the applications of statistical analysis to their research problems. It is my hope that this book will help in providing the tools to make that possible.

HOWARD B. STAUFFER

Mathematics Department
Humboldt State University
Arcata, California

1 Introduction

We will begin this initial chapter by introducing three case studies that illustrate some of the fundamental general statistical problems challenging the contemporary natural resource scientist. We will then present a review and preview of some solution strategies to these general problems. The first solution strategies that we will review are traditional frequentist approaches: parameter estimation from sample surveys, hypothesis testing from experiments, and linear regression modeling. Each of these methods is summarized using a frequentist approach to statistical analysis. We will then preview some more contemporary solution strategies: an alternative Bayesian approach to statistical analysis and other more advanced solutions to the case studies, generalized linear modeling, and mixed-effects modeling using both frequentist and Bayesian approaches to statistical analysis. We will also preview a more contemporary approach to model selection and inference using information-theoretic criteria such as Akaike's information criterion for frequentist statistical analysis and the deviance information criterion for Bayesian statistical analysis. All of these contemporary methods will be discussed in greater detail throughout the remainder of this book and illustrated with examples.

In this initial chapter we include a reminder of the importance of project management in natural resource studies with statistical components. Project management consists of organizing projects into three phases: a planning phase, a data collection phase, and a concluding phase. The planning phase includes an identification of the problem and the objectives of the project, along with a statistical design for the collection of the dataset. The concluding phase includes a statistical analysis of the dataset, along with interpretation and conclusions drawn from the analysis. All of these statistical components—the statistical design, the collection of the dataset, and the statistical analysis—provide essential tools for the solutions to the objectives of the project.

We conclude this initial chapter with an introduction to the frequentist statistical analysis software used throughout the book: the proprietary software S-Plus and its freeware "equivalent" R. The Bayesian statistical analysis software WinBUGS will be introduced in Chapters 2–4 when Bayesian ideas are discussed.

Contemporary Bayesian and Frequentist Statistical Research Methods for Natural Resource Scientists. By Howard B. Stauffer

1.1 INTRODUCTION

In recent years there have been major advances in the methods of statistics used for research in the natural resource sciences. Yet, little of this is known outside selected research circles. Students and scientists in the natural resource sciences have continued to use traditional frequentist methods, such as the estimation of parameters from sample surveys, t tests and ANOVA hypothesis testing from experiments, and linear regression modeling. However, extraordinary newer methods are now available that enhance, complement, and extend these basic techniques, methods such as Bayesian statistical inference, information-theoretic approaches to model selection, generalized linear modeling, and mixed-effects modeling. It is the primary objective of this book to introduce these newer contemporary methods to natural resource students and scientists.

This book must begin by emphasizing critical statistical issues that have too often been neglected in natural resource studies in the past. We stress the importance of the planning and concluding phases in a data collection project. We particularly highlight the essential role of statistical design and analysis that help ensure the efficient, powerful, and effective use of data. Our approach throughout the book will be "hands-on," illustrating concepts with examples using the software languages of S-Plus or R for frequentist statistical analysis and WinBUGS for Bayesian statistical analysis.

Let's begin with a description of several case studies that illustrate problems of fundamental interest to contemporary natural resource scientists.

1.2 THREE CASE STUDIES

1.2.1 Case Study 1: Maintenance of a Population Parameter Above a Critical Threshold Level

A fundamental problem of interest to contemporary natural resource scientists is to assess whether a critical population parameter, such as a proportion parameter p, has been maintained above (or below) a specified critical threshold level: $p \geq p_c$ (or $p \leq p_c$)?

Many examples in natural resource science illustrate this problem:

1. A timber company is required to maintain the proportion p of its timberlands occupied by nesting Northern Spotted Owl pairs above a specified threshold level p_c. The threshold p_c is a level determined by biologists to ensure the viability of the local population of owls.
2. Federal managers of a national forest are interested in maintaining the proportion p of forest covered by dense undergrowth below a specified threshold level p_c, to limit the risk of fire.
3. The managers of a national park are interested in maintaining the proportion p of a disease or insect infestation below a specified threshold level p_c to control its spread.

4. Fishery biologists managing a watershed are interested in maintaining the proportional abundance p of a fishery above a specified threshold level p_c of its carrying capacity to ensure its long-term sustainability.
5. A government agency implementing a natural resource conservation policy is interested in ensuring that the proportion p of the public in favor of one of its controversial policies is maintained above a certain threshold level p_c.

Besides the proportion parameter p in the examples presented above, there are many other biological parameters of interest to natural resource managers with similar threshold issues, such as the mean abundance μ, survival rate ϕ_i from year i to year $i + 1$, fitness $\lambda_i = N_{i+1}/N_i$ (where N_i and N_{i+1} are the population abundances in years i and $i + 1$), ecological diversity index such as the Shannon–Wiener diversity index H, and population total τ.

The failure to maintain the population parameter p above (or below) the threshold level p_c might suggest the need for a "corrective action" decision in the examples listed above, such as

1. Reducing the timber harvesting
2. Applying fire suppression treatment
3. Applying disease or insect treatment
4. Increasing the watershed river flow by releasing more water from a dam
5. Altering the natural resource conservation policy

Alternatively, success at maintaining the population parameter p above (or below) the threshold p_c might suggest a decision of "no action."

In such circumstances, a common approach employed by natural resource scientists is to begin monitoring the population and collecting sample data, say, on an annual basis, in order to assess the status of the population parameter. The intent is to conduct statistical analysis on the sample data and make inferences about the population parameter to determine whether it is above (or below) the threshold, and thus whether corrective action or no action is needed at the management level.

1.2.2 Case Study 2: Estimation of the Abundance of a Discrete Population

Our second case study focuses on the analysis of population count data. Often biological populations, such as birds, amphibians, or mammals, are sampled with discrete measurements such as plot counts, in fixed-area plots called quadrants. The intent is to estimate population size or density in an area using total count estimates of abundance or mean estimates of density.

The analysis consists of estimates of total or mean. Traditional estimates of total or mean are based on the assumption of the normal distribution of the population measurements. For plot counts, however, measurements are discrete and noncontinuous, consisting of nonnegative integers in a skewed distribution. If the biological

population is randomly dispersed spatially, a proper model for the analysis should be based on the Poisson distribution rather than the normal distribution for the plot counts. If, however, the population is spatially aggregated or clumped, the analysis should be based on a more general model for the population measurements, such as a negative binomial distribution. Furthermore, the plot counts will likely be sampled without complete certainty of detection. Animals may be within the plot and yet be undetected by the sample surveyor. A rigorous analysis of the population therefore must factor in the Poisson or negative binomial distribution of the plot measurements, sampled with an uncertainty of detection. We will examine such analyses in Chapters 2–4 with Bayesian statistical analysis and Chapters 6–7 with generalized linear models and mixed-effects models.

1.2.3 Case Study 3: Habitat Selection Modeling of a Wildlife Population

In general, it can be quite difficult to estimate the presence or abundance of a wildlife population. Many important biological populations whose presence or abundance needs to be estimated are endangered or locally threatened wildlife species, such as the Northern Spotted Owl and Marbled Murrelet bird populations, Del Norte salamander amphibian population, and grizzly bear mammal population. These endangered species are often of particular importance because they are associated with old-growth ecosystems that are also in danger of extinction. Therefore it is important to monitor these populations, estimating their presence or abundance over time, to assess the status of the old-growth ecosystems. A particularly effective approach to estimating these mobile populations is to model their relationship with habitat.

With **habitat selection modeling**, the presence or abundance of a mobile population species is treated as a dependent response variable. Its relationship with "independent" predictor explanatory habitat variables such as vegetation, geologic, and meteorologic attributes can be assessed with statistical modeling. The intent of the habitat selection modeling is to analyze the relationship between the mobile wildlife population variable and the habitat variables and use it to describe or predict the presence or abundance of the endangered species as a function of the habitat variables. The idea behind the modeling is that many habitat variables can be more easily and less expensively sampled than can the mobile wildlife population.

The relationship in such circumstances is assumed to be associative rather than causal; thus, the modeling is descriptive, based on population monitoring with sample survey data, and not on experimental manipulation to establish evidence for cause and effect. The mobile wildlife population may have access to only a limited amount of habitat attributes and be able to express a restricted preference among what remains. Other habitat attributes that the mobile wildlife population most prefers may no longer be available for selection. Hence the habitat "selection" relationship must be interpreted within this context.

Habitat selection modeling is often based on regression analysis. For continuous-abundance response variables such as biomass, multiple linear regression analysis may indeed be applicable. For discrete-abundance response variables such as population counts, however, Poisson regression or negative binomial regression may be

more appropriate. For binary response variables, such as the presence or absence, or occupancy versus nonoccupancy, of a population, logistic regression analysis or some other form of generalized linear modeling may be more appropriate. We will examine these methods of analysis, along with strategies for model selection, in Chapters 5 and 6. Traditional multiple linear regression analysis is discussed in Chapter 5. Logistic regression analysis, Poisson regression analysis, negative binomial regression analysis, and other forms of generalized linear modeling are discussed in Chapter 6.

1.2.4 Case Studies Summary

This book presents various contemporary statistical options available to the natural resource scientist to analyze and interpret sample data for these case studies and other general statistical problems of current interest to natural resource scientists. We will first review more familiar traditional statistical methods of sample survey parameter estimation, experimental hypothesis testing, and multiple linear regression modeling, and then describe the less familiar contemporary methods of Bayesian statistical inference, model selection strategies, generalized linear modeling, and mixed-effects modeling. These methods provide contemporary natural resource scientists with an up-to-date statistical toolbox of methods to tackle many important challenging problems of current interest.

1.3 OVERVIEW OF SOME SOLUTION STRATEGIES

In this section we present both a review of traditional statistical methods and a preview of contemporary statistical methods that provide solutions to the case studies that were presented in the previous section: assessing whether a population parameter has been maintained above (or below) a critical threshold level, the estimation of abundance of a discrete population, and habitat selection modeling. Further details on the contemporary methods will follow in later chapters.

1.3.1 Sample Surveys and Parameter Estimation

A first traditional statistical approach to addressing the fundamental case study problems of Section 1.2 is to conduct a sample survey of the population and collect sample data using a rigorous sampling design. The aim of the survey in case study 1 is to estimate a proportion parameter p or mean parameter μ from the sample data and compare it with a critical threshold level p_c or μ_c. The aim of the survey in case study 2 is to estimate the mean abundance parameter μ of a discrete population. The aim of the survey in case study 3 is to model a mobile wildlife population as a function of habitat attributes and estimate the proportion parameter p or abundance parameter mean μ of the species in the habitat or at a specific site. Ideally, a natural resource scientist would like to use an approximately unbiased **estimator** $\hat{\theta} = \hat{p}$ or $\hat{\mu}$ for the **estimate** of the **parameter** $\theta = p$ or μ, respectively, of minimum **sampling**

error E, with a specified **level of confidence** P (or **level of significance** $\alpha = 1 - P$).

If simple randomly sampled measurements $\{y_i\}$ are continuous and normally distributed with sample size n, the **mean estimator** is given by

$$\hat{\mu} = \sum_{i=1}^{n} \frac{y_i}{n},$$

with **standard deviation**

$$s = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu})^2}{n-1}},$$

standard error

$$\text{se} = \frac{s}{\sqrt{n}},$$

and **sampling error**

$$E = t_{(1-\alpha/2),n-1} \cdot \text{se} = t_{(1-\alpha/2),n-1} \cdot \frac{s}{\sqrt{n}},$$

where $t_{(1-\alpha/2),\,n-1}$ is the t value with $(n-1)$ degrees of freedom at the $(1-\alpha/2)$ percentile with α level of significance.

If simple randomly sampled measurements $\{y_i\}$ are binary and binomially distributed with sample size $n \geq 30$, the **proportion estimator** is given by

$$\hat{p} = \frac{y}{n}, \quad \text{where} \quad y = \sum_{i=1}^{n} y_i,$$

with standard error

$$\text{se} = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n-1}},$$

and sampling error

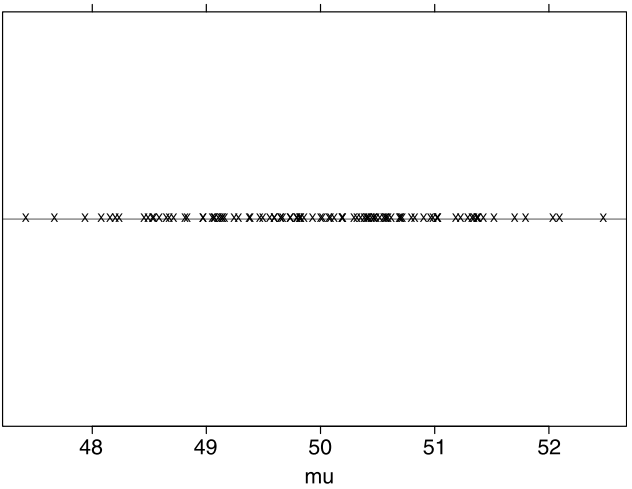
$$E = t_{(1-\alpha/2),n-1} \cdot \text{se},$$

where $t_{(1-\alpha/2),n-1}$ is the t value with $(n-1)$ degrees of freedom at the $(1-\alpha/2)$ percentile with α level of significance.

Recall that an **unbiased estimator** has the property that the average of all estimates, with repeated sampling, is equal to the parameter value (Fig. 1.1). Confidence levels of $P = 95\%$, 90%, or 80% are commonly used for natural resource survey sampling with levels of significance $\alpha = 1 - P = 5\%$, 10%, and 20%, respectively. A **confidence interval**

$$\text{CI} = [\hat{\theta} - E, \hat{\theta} + E]$$

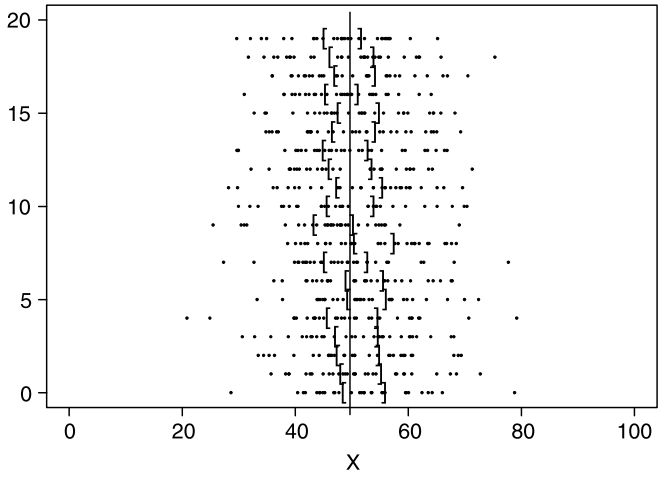
Figure 1.1. Sample mean density estimates (tickmarks X), based on 100 repeated sample surveys of a normally distributed population with mean density parameter value 50.0. Note that the average of these sample mean density estimates is approximately equal to the mean density parameter value for this unbiased mean estimator.



can be calculated and the frequentist inference drawn that there is a probability P that confidence intervals will contain the parameter θ, with repeated sampling (Fig. 1.2).

Note that the logic of frequentist inference is of the form "if (parameter), then probability(data)." It assumes that the parameter is fixed and provides conditional probability properties for statistics from the sample datasets.

Figure 1.2. Twenty 95% confidence intervals estimated from samples obtained from repeated sample surveys of a population with mean parameter value 50.0 ("|"). Note that 19 of these confidence intervals contain the mean parameter, as is expected.



For case study 1, a **decision protocol** should be specified in advance, before the sample data are collected. For example, one such decision protocol would be to compare the estimate $\hat{\theta}$ obtained from the sample data with the critical threshold level θ_c. If the estimate is above (or below) the critical threshold level θ_c, then the recommended management decision would be "corrective action." Otherwise, if the estimate is below (or above) θ_c, the recommended management decision would be "no action."

An alternative decision protocol for case study 1 would be to compare the confidence interval CI with the critical threshold level. If the confidence interval is above (or below) the critical threshold level, then the recommended management decision would be "corrective action." If the confidence interval is below (or above) the critical threshold level, the recommended management decision would be "no action." If the confidence interval overlaps the critical threshold level, the situation would be ambiguous and need to be reassessed, perhaps with an additional survey with larger sample size. The **precision** of the estimate, the size of the sampling error, would obviously affect the results. A larger sample size would reduce the sampling error and hence the size of the confidence interval. Therefore, the population should be sampled with a sample size large enough to reduce the sampling error so that the confidence interval will (hopefully!) fall on one side or the other of the critical threshold level.

Other decision protocols could be chosen for case study 1 using this general approach of sample surveys with parameter estimation and estimation of error. The important point, however, is that a decision protocol for a sample survey should be specified in advance of data collection so that a decision can be made, clearly and unambiguously, at the end of the survey. In Chapters 2–4 we shall see how a Bayesian statistical analysis approach can facilitate the use of a decision protocol.

For case studies 2 and 3, estimates and confidence intervals can be used to make inferences on the abundance of a discrete population, and the population habitat selection probability of presence or mean abundance, respectively. For further review of the basic concepts of sampling design and analysis, see Cochran (1977), Scheaffer et al. (1996), Thompson (1992), Sarndal et al. (1992), Thompson and Seber (1996), Thompson et al. (1998), Stauffer (1982a, 1982b), Hansen and Hurwitz (1943), Horvitz and Thompson (1952), and Gregoire (1998).

1.3.2 Experiments and Hypothesis Testing

A second statistical approach to addressing the problems posed by some of the case studies of Section 1.2 is to conduct an experiment and use frequentist hypothesis testing, developed by Neyman and Pearson (1928a, 1928b, 1933, 1936) and Fisher (1922, 1925a, 1925b, 1934, 1958). With case study 1, for example, the scientist could formulate the null hypothesis

$$H_0 : p = p_c$$

and the one-tailed alternative hypothesis

$$H_A : p < p_c,$$

collect experimental data using a rigorous experimental design, and test the null hypothesis. If the null hypothesis is rejected, the recommended management decision would be "corrective action." If the null hypothesis is not rejected, the recommended management decision would be "no action." Note, that with this approach, the burden of proof would be on "corrective action." If the direction of the alternative hypothesis is reversed, the burden of proof would be on "no action." Regardless, the burden of proof would not be equal for the two hypotheses.

A one-sample z test could be used for the hypothesis testing, with a one-sided alternative hypothesis. If the null hypothesis is true, the test statistic

$$z_s = \frac{\hat{p} - p_c}{\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n - 1}}}$$

is standard normally distributed. The **Neyman–Pearson hypothesis testing protocol** requires that a **type I error** α be specified, with confidence $P = 1 - \alpha$ (say, $\alpha = 5\%$ and $P = 95\%$), prior to the data collection and analysis, assuming the null hypothesis to be true. If the test statistic z_s, calculated from the experimental dataset, is in the rejection region, the α percentile left tail of the standard normal distribution (equivalent to $p < \alpha$), then the null hypothesis would be rejected. Otherwise, the null hypothesis would not be rejected.

Note again, that the logic of frequentist inference is of the form "if (parameter), then probability (data)." It assumes the null hypothesis that the parameter is fixed and provides conditional probability properties for statistics from the experimental datasets.

To review the basic concepts of experimental design and hypothesis testing, see Hicks (1993), Kuehl (1994), Dowdy and Wearden (1991), Sokal and Rohlf (1995), Zar (1996), Winer et al. (1991), Cohen (1988), Siegel and Castellan (1988), Conover (1980), Daniel (1990), PASS (2002), and nQuery (2002). However, beware of the overuse and misuse of hypothesis testing in natural resource science, particularly with observational, rather than experimental, datasets; see Johnson (1999), Anderson et al. (2001, 2002), Ramsey and Schafer (2002), and Robinson and Wainer (2002).

1.3.3 Multiple Linear Regression, Generalized Linear Modeling, and Model Selection

Habitat selection modeling can often be used effectively to address case study 3 of Section 1.2. Mobile wildlife populations may be difficult to sample directly, but their responses often vary with habitat attributes that are more easily sampled. For instance, the response of endangered wildlife species associated with old-growth habitat may tend to be larger in value in late-seral-stage rather than early-seral-stage habitat. In such circumstances, it may be useful, more cost-effective, and less time-consuming to fit and compare habitat selection models that describe the mobile biological population response as a function of various habitat variables.

The objective is to express the wildlife population response as a function of variables describing the vegetation, geologic, and climatic characteristics of its habitat. If the population response is continuous with normally distributed error, the population may be described by a multiple linear regression model.

If, however, the population response measurements are categorical, such as binary with values of 0 or 1 (e.g., "present" or "absent," "occupied" or "unoccupied," "alive" or "dead," "yes" or "no"), or discrete with integer values such as plot counts with error that may not be normally distributed, then it may be possible to "link" the response measurements to a linear function described by a generalized linear model (GLM) such as logistic regression.

Multiple linear regression models and generalized linear models can be used to describe the response values at specific sites as a function of habitat characteristics. With such a modeling approach to statistical analysis, model selection strategies are required to effectively compare models for goodness of fit to sample datasets and to avoid overfitting and compounding of error. The utility of such an approach depends on the goodness of fit of the best-fitting models to the population and their predictive accuracy. Chapter 5 includes a basic review of multiple linear regression and a description of contemporary strategies that can be used for model selection and inference. To further review the details of multiple linear regression, see Appendix A at the end of this book, or see Seber (1977), Draper and Smith (1981), Hocking (1996), Ryan (1997), Cook (1998), and Cook and Weisberg (1999). Generalized linear modeling is introduced in Chapter 6.

1.3.4 A Preview of Bayesian Statistical Inference

Bayesian statistical analysis and inference provides an important alternative approach to frequentist statistical analysis and inference for natural resource scientists, yet it has become practical and accessible for general use only relatively recently. Traditional frequentist statistical analysis and inference provides probabilities for sample datasets, based on assumptions for parameters, with an interpretation of results in the context of repeated surveys or experiments. Although frequentist statistical analysis methods are well known by natural resource scientists, these methods are often incorrectly applied with inferences that are frequently misunderstood (Johnson 1999, 2002). If properly applied and correctly interpreted, frequentist statistical analysis provides rigorous standards for inferences: the unbiased and minimum error properties of estimators, the accuracy probabilities of confidence intervals, and the type I and type II errors of hypothesis testing.

Alternatively, Bayesian statistical analysis and inference provides probabilities for parameters, based on sample datasets (Iversen 1984, Berger 1985). Inferences from Bayesian statistical analysis are directly applicable to parameters that are of central interest to natural resource scientists. Unfortunately, Bayesian statistical inference has not been of leading interest to a majority of statisticians and practicing natural resource scientists in the past because its use has been impractical and inaccessible until very recently (as of 2007). Bayesian statistical analysis and inference requires an assumption of a prior distribution for the parameters. Using the sample dataset

and a likelihood model for the dataset, Bayesian statistical analysis provides a posterior distribution for the parameters, based on the prior distribution, the dataset, and the model. The posterior distribution thus updates a scientist's understanding of the parameters. The Bayesian approach combines previous information about the parameter with an analysis of the sample dataset to obtain an updated assessment of the parameters. The posterior distribution provides probabilities for the parameters that can be useful for natural resource scientists and managers. They can utilize summary statistics of the posterior distribution, such as the mean, median, mode, standard deviation, and percentiles, or use the entire posterior distribution itself to evaluate the parameters. A probability region, the smallest middle interval encompassing 95% of the posterior distribution, provides a Bayesian 95% **credible interval**. This interval can be directly interpreted as the region within which the parameter is likely to be found, with 95% probability. Thus, with Bayesian statistical inference, there is no need to interpret the results indirectly as a frequentist does, in terms of probabilities of datasets with repeated surveys or experiments. The logic of Bayesian statistical analysis provides probabilities for parameters, given the data, in contrast to the logic of frequentist statistical analysis, which provides probabilities for datasets, given the parameter.

Bayesians must, however, bear responsibility for the appropriate selection of priors and the standards of results. As with frequentist results, Bayesian results must be assessed for goodness of fit to assess the reliability of model predictions. Priors influence posteriors, particularly with small-sample datasets, and must be chosen judiciously. Until relatively recently, Bayesian solutions to complex problems were seldom computable. However, owing to a collection of computer simulation algorithms, the Markov Chain Monte Carlo (MCMC) algorithms developed in the mid-twentieth century (Bremaud 1999, Carlin and Louis 2000, Congdon 2001, Gill 2002, Link et al. 2002) and to public-domain software such as WinBUGS that is now downloadable from the Web (Spiegelhalter et al. 2001), natural resource scientists now have the resources needed for the practical use of Bayesian statistical inference.

This book describes and illustrates both frequentist and Bayesian paradigms for statistical analysis and inference, emphasizing the advantages and disadvantages of each in particular contexts. It is the practical perspective of this book that the contemporary natural resource scientist should be familiar with both. Bayesian statistical inference is introduced in Chapters 2–4 and applied comparatively, along with frequentist statistical inference, in Chapters 5–7, with other important contemporary research methods of statistical analysis.

1.3.5 A Preview of Model Selection Strategies and Information-Theoretic Criteria for Model Selection

With either frequentist or Bayesian statistical analysis approaches, a rigorous and theoretically justifiable approach to model fitting, selection, and inference is required. Traditionally, with multiple linear regression modeling, analysts have used statistics such as parameter coefficient estimates and their significance, the coefficient of

determination R^2, the residual standard error $s_y|_x$, the ANOVA F test, the adjusted R^2, and Mallows' C_p to evaluate the relative fit of models (Seber 1977, Draper and Smith 1981, Hocking 1996, Ryan 1997, Cook and Weisberg 1999). These statistics test various assumptions of the model fit, such as whether the model is statistically equivalent to the null model. They do not directly assess the issue of whether the model is the best fitting to the sample dataset. Unfortunately, they also sometimes tend to overfit the model to the sample dataset, with compounding of error. We shall say more about this later on. We recommend a more modern information-theoretic approach to model fitting, using Akaike's information criterion (AIC), the corrected Akaike information criterion (AIC_c), or the Bayesian information criterion (BIC) with frequentist statistical analysis (Burnham and Anderson 1998, 2002), and the deviance information criterion (DIC) with Bayesian statistical analysis (Spiegelhalter et al. 2001, Carlin and Louis 2000). These criteria provide a more rigorous and theoretically justified approach to model fitting that avoids the overfitting of models to the sample dataset and the compounding of error.

Akaike's information criterion was developed relatively recently by the Japanese mathematician Hirotugu Akaike (1973, 1974). It is an information-theoretic measurement of the relative **Kullback–Leibler distance (KL distance)** between a model and the reality. The **Akaike's information criterion (AIC)** is the linear Taylor series approximation of the relative KL distance, whereas the **corrected Akaike information criterion (AIC_c)** is a second-order Taylor series approximation. Since AIC_c is more precise, we recommend that it be used in preference to AIC, particularly for datasets with small numbers of samples. The best-fitting model in a collection of models has the lowest AIC or AIC_c value.

For any probabilistic statistical model with a likelihood function $\mathcal{L}$ (more on this in Chapters 2, 5, and 6), AIC and AIC_c are defined using the deviance $= D = -2 \cdot \log(\mathcal{L}) = -2 \cdot l$

$$\begin{aligned} \text{AIC} &= D + 2 \cdot k \\ &= -2 \cdot \log(\mathcal{L}) + 2 \cdot k \end{aligned}$$

and

$$\begin{aligned} \text{AIC}_\text{c} &= D + 2 \cdot k + 2 \cdot \frac{k \cdot (k+1)}{n-k-1} \\ &= -2 \cdot \log(\mathcal{L}) + 2 \cdot k + 2 \cdot \frac{k \cdot (k+1)}{n-k-1}, \end{aligned}$$

where $k =$ the number of parameters in the model and $n =$ the sample size.

The AIC and AIC_c criteria for multiple linear regression are given by the formulas

$$\begin{aligned} \text{AIC} &= n \cdot \log\left(\hat{\sigma}^2 \cdot \frac{n-p-1}{n}\right) + 2 \cdot k \\ &= n \cdot \log\left(\hat{\sigma}^2 \cdot \frac{n-k+1}{n}\right) + 2 \cdot k \end{aligned}$$

and

$$\begin{aligned}\mathrm{AIC_c} &= n \cdot \log\left(\hat{\sigma}^2 \cdot \frac{n-p-1}{n}\right) + 2 \cdot k + 2 \cdot \frac{k \cdot (k+1)}{n-k-1} \\ &= n \cdot \log\left(\hat{\sigma}^2 \cdot \frac{n-k+1}{n}\right) + 2 \cdot k + 2 \cdot \frac{k \cdot (k+1)}{n-k-1},\end{aligned}$$

where $\hat{\sigma} = s_{y|x}$ is the residual standard error, p is the number of covariates or explanatory variables, and $k = p + 2$ is the number of parameters (including the covariates coefficients, the intercept, and σ). For the linear regression model with the parameters β_0, β_1, and σ, $p = 1$ and $k = 3$.

The $\mathrm{AIC_c}$ criterion penalizes a model with too many covariates from overfitting the sample data. It determines the most parsimonious model, the one with an optimum mix of minimal bias and maximal precision. As the number of parameters increases, models more closely fit sample datasets, reducing the bias. However, as the number of parameters per sample increases, the precision of the parameter estimates tends to decrease. The $\mathrm{AIC_c}$ criterion moderates this process, striking the most optimal compromise between reduced bias and maximal precision.

The corrected Akaike information criterion measures the amount of noise, or **entropy**, in the sample data, separating it from the **signal** or **information**. It is a relative measure of the KL distance between the model and the reality. The absolute measure of the entropy is the calculated $\mathrm{AIC_c}$ plus a constant. The constant remains unknown since the reality is unknown. However, since each model has the same constant, $\mathrm{AIC_c}$s may be compared to determine the relatively best-fitting model. The reader should be warned, however, that this fit is relative. Goodness-of-fit tests must additionally be used in the concluding analysis to assess the absolute fit of the best-fitting models with the lowest $\mathrm{AIC_c}$s.

The $\mathrm{AIC_c}$ is most applicable to models of realities that are complex and infinite- or high-dimensional, as are most natural resource populations. For such complex realities, finite-dimensional models will necessarily be at best only an approximation. For realities that are finite-dimensional, of fairly low dimension, such as $k = 1$–5, with k fixed as the sample size n increases, "dimension-consistent" criteria such as the Bayesian information criterion are more applicable (Burnham and Anderson 1998, 2002).

The **Bayesian information criterion (BIC)**, developed by Schwarz (1978), also uses a formula based on the deviance or log likelihood and "penalizes" models for the overuse of covariates

$$\begin{aligned}\mathrm{BIC} &= D + k \cdot \log(n) \\ &= -2 \cdot \log(\mathcal{L}) + k \cdot \log(n).\end{aligned}$$

For multiple linear regression models, BIC is given by

$$\begin{aligned}\mathrm{BIC} &= n \cdot \log\left(\hat{\sigma}^2 \cdot \frac{n-p-1}{n}\right) + k \cdot \log(n) \\ &= n \cdot \log\left(\hat{\sigma}^2 \cdot \frac{n-k+1}{n}\right) + k \cdot \log(n).\end{aligned}$$

The BIC is derived using Bayesian assumptions of equal priors for each model and vague priors on the parameters (Burnham and Anderson 1998), with the objective of predicting rather than understanding the process of a system. The BIC penalizes more heavily for increases in the number of parameters and hence sometimes tends to select models that are underfit with excessive bias and precision. For natural resource modeling, most realities are complex and infinite-dimensional; hence, AIC_c is the more appropriate criterion for comparing statistical models in the natural resource sciences.

1.3.6 A Preview of Mixed-Effects Modeling

Data collected from monitoring do not always fulfill the assumptions of independence required for the use of many traditional statistical methods: with parameter estimation in sample surveys, with hypothesis testing in experiments, and with model fitting and selection in multiple linear regression or generalized linear modeling. Rather, data are often dependent, clustered or grouped by location or time, or collected from permanent plots or at sites repeatedly over time or from subpopulations of larger populations (e.g., as with meta-population data). Traditionally, scientists have often been discouraged from collecting dependent or pseudoreplicated datasets to avoid these problems with the analysis. But dependencies in biological populations are quite common, and it may be difficult or impossible to avoid collecting data with such dependencies. It would make more sense to collect data with such dependencies and account for them in the analysis. This is now possible with **mixed-effects modeling**, which incorporates both traditional **fixed effects** describing the influences of **treatments** on the population and **random effects** describing dependencies in the data created by groupings. This is achieved in an efficient manner with mixed-effects modeling, incorporating variance components into the models to describe the random effects due to the clusters. We will provide an introduction to mixed-effects modeling using the powerful utilities now available in S-Plus and R for frequentist analysis and WinBUGS for Bayesian analysis in Chapter 7.

1.4 REVIEW: PRINCIPLES OF PROJECT MANAGEMENT

In this section of this introductory chapter, we remind the reader of the critical importance in a natural resource data collection project of practicing the principles of sound project management. A data collection project consists of three phases: a planning phase at the beginning, a data collection phase in the middle, and a concluding phase at the end.

It is very important at the beginning of a project to devote sufficient attention to the planning phase in order to develop a rigorous and effective statistical design for the data collection. To properly determine the appropriate statistical design, the problem, objectives, and methods of analysis for the project must be clearly specified prior to data collection. Far too often in natural resource data collection projects, the problem and objectives are not specified clearly enough in quantitative terms. If a project, for example, has the objective of examining the downward trend of a declining species,

the amount of downward trend that is biologically important should be specified in advance of data collection. Then the sample size for the data collection should be calculated to ensure with a high probability that the biologically significant trend will be detected with statistical significance if it exists.

It is unfortunately also far too common in data collection projects to wait until after the data have been collected to decide on the method of analysis. As the method of analysis required by the objectives may impose restrictions on the statistical design of the data collection, it should be determined prior to data collection. The method of analysis may require a certain amount of precision or power to realize the objectives of the project, and this will require a sufficiently large sample or experimental dataset. Hence, the sample size or numbers of replicates must be determined prior to data collection.

It is also vitally important to devote sufficient attention to the concluding phase at the end of a project, to allow time for a comprehensive analysis and thoughtful interpretation and conclusions. Comprehensive analysis will seldom be "turnkey," which can be finished in a few hours or days, but rather a far more lengthy process. The analysis process may consist of examining a range of candidate models to determine the best fitting. This collection of models may be small and finite or wide-ranging in number. We shall describe alternative strategies for the model selection and inference in Chapter 6.

It is tempting in a natural resource data collection project to devote a majority of the time to data collection. This unfortunately may leave an insufficient amount of time at the beginning of the project for planning and at the end of the project for a thorough analysis and thoughtful period of reflection on the conclusions and interpretation of the results. A good principle in general is to spend equal amounts of effort on all phases of the project. This will help ensure that the results are efficient, powerful, and effective. The aim is to extract the biological "information" from the data, separating the "signal" from the "noise" in as optimal a manner as possible. Time and effort devoted to rigorous design and analysis in a data collection project are indeed well spent.

1.5 APPLICATIONS

The practical application of theory is where the learning process really crystallizes. This is particularly true with statistical analysis. We will use a hands-on approach to emphasize the practical use of statistics with the application of theory. We will use both simulated and real-world sample datasets as examples and encourage the reader to do likewise with other datasets.

We encourage readers to analyze their own datasets while progressing through the book. Readers are especially encouraged to conduct a project while reading this book, designing, collecting, and analyzing their own sample or experimental datasets. Address a biologically interesting question, but one sufficiently limited in scope that it can be investigated with a data collection project completed within the timeframe of the reading of this book. For example, examine the abundance level or trend of a

local biological population of interest such as a bird or plant population. Write a 2–5-page proposal for the project. Keep the project small and realistically simple; be especially careful to keep the scope of the project within a realistic timeframe.

We will use simulated datasets to illustrate many of the ideas. The reader will thus be able to compare the statistical results with the known "realities" used to generate the simulated datasets. We will also provide real-world datasets for the analyses of important species such as the Northern Spotted Owl, Siskiyou Mountains salamander, and beach layia, some of which are endangered. These datasets will provide practical, realistic experience. The interested reader can also obtain additional sample and experimental datasets on the Web and in many other standard references, such as Sokal and Rohlf (1995), Zar (1996), and Ramsey and Schafer (2002). These references contain excellent examples, problems, and datasets.

1.6 S-Plus® AND R ORIENTATION I: INTRODUCTION

1.6.1 Orientation I

In this section, we provide an introductory orientation to the statistical software used for the frequentist analysis throughout the book, the research-oriented proprietary statistical modeling software S-Plus® (2000) and the de facto equivalent freeware R (R Development Core Team 2005). S-Plus was first developed in the 1970s at Bell Labs by Rick Becker, John Chambers, and Allan Wilks with the goal of defining a language to perform repetitive tasks in data analysis. These authors did not consider the original language to be primarily statistical and most of the statistical functionality was added later. S-Plus provides a flexible, interactive, integrated modern environment for statistical analysis, with particular emphasis on the modeling of linear systems. Much of the syntax of S-Plus is reminiscent of the Unix and C environments.

The current version of R is the result of a collaborative effort with contributions from all over the world. The R language was initially written by Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland. It is freeware, the de facto equivalent of S-Plus.

Details on S-Plus can be found in the S-Plus `Help` menu, the S-Plus Manual (S-Plus 2000), or Krause and Olson (2000). Details on R can be found in the R `Help` menu. S-Plus and R are object-oriented languages with datasets and functions consisting of objects. Objects can be created and manipulated by the user and added to the standard library of defaulted objects that are available in S-Plus and R. To begin this introductory S-Plus–R session, the user should sign onto either S-Plus or R. In S-Plus, the user should enter the `Commands Window` mode from the `Window` menu to begin at the > prompt. In R, the user should begin at the > prompt in the R Console. Although S-Plus and R do have some menu features, the S-Plus command mode and R Console options will be emphasized throughout this book, leaving it to readers to explore the menu options. We will present frequentist software code that is sufficiently general for use in either S-Plus or R and will be sufficiently careful to point out where there are differences.

1.6.2 Simple Manipulations

Let's begin by illustrating simple data manipulation capabilities in S-Plus and R, starting with arithmetic using +, −, *, /, and ∧ for addition, subtraction, multiplication, division, and exponentiation, respectively. Proceed with the following code (Fig. 1.3a), typing the first line, pressing the <ENTER> key, receiving the second line response from S-Plus or R, typing the third line and pressing the <ENTER> key to receive the fourth-line response, and so on. The index [1] at the beginning of each response line refers to the first indexed entry of the arithmetic operation response, a vector of size 1.

For the next illustration, the c concatenation, rep repetition, 1:k integer sequence, and seq general sequence commands aid in constructing data objects in

Figure 1.3. Command code for S-Plus and R Orientation I. (**a**) Simple manipulation: arithmetic; (**b**) Simple manipulations: creation of vectors; (**c**) Simple manipulations: object removal; (**d**) Data structures: vectors, data frames, lists; (**e**) Data structures: testing; (**f**) Data structures: coercion; (**g**) Random numbers: normal distribution; (**h**) Random numbers: sampling; (**i**) Directory structure; (**j**) Functions and control structures: functions code. (**k**) Functions and control structures: functions execution; (**l**) Linear regression analysis.

(**a**)

```
> 9 + 3 <ENTER>
[1] 12
> 9 - 3 <ENTER>
[1] 6
> 9 * 3 <ENTER>
[1] 27
> 9 / 3 <ENTER>
[1] 3
> 9 ^ 3 <ENTER>
[1] 729
```

(**b**)

```
> c(2,5,4,9)
[1] 2 5 4 9
> rep(5,4)
[1] 5 5 5 5
> 2:9
[1] 2 3 4 5 6 7 8 9
> seq(0,1,.1)
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

(**c**)

```
> x <- c(2,4)
> x
[1] 2 4
> rm(x)
> x
Error: Object "x" not found
```

Figure 1.3. *Continued.*

(**d**)

```
> v1 <- c(2,3,5)
> v1
[1] 2 3 5
> v2 <- c(5,3,7)
> v2[3]
[1] 7
> v3 <- c("low","medium","high")
> v4 <- c(T,F,T,T)
> m1 <- cbind(v1,v2)
> m1
     v1 v2
[1,]  2  5
[2,]  3  3
[3,]  5  7
> d1 <- data.frame(v1,v3)
> d1
  v1     v3
1  2    low
2  3 medium
3  5   high
> list1 <- list(v1,v4)
> list1
[[1]]:
[1] 2 3 5
[[2]]:
[1] T F T T
```

(**e**)

```
> is.numeric(v1)
[1] T
> is.numeric(v3)
[1] F
```

(**f**)

```
> v1
[1] 2 3 5
> as.character(v1)
[1] "2" "3" "5"
> is.numeric(as.character(v1))
[1] F
```

Figure 1.3. *Continued.*

(**g**)

```
> y <- rnorm(10,5,1)
> y
 [1] 6.444356 3.777367 3.974797 5.568687 5.875846
 [6] 3.441874 5.037888 7.006315 4.543586 2.957072
> pnorm(4,5,1)
[1] 0.1586553
> qnorm(.1586553,5,1)
[1] 4
> dnorm(5,5,1)
[1] 0.3989423
```

(**h**)

```
> x <- 1:10
> x
 [1]  1  2  3  4  5  6  7  8  9 10
> sample(x,5)
[1] 3 8 1 9 7
> sample(x,15)
Error in sample(x,15): Population not large enough
for given sample size
> sample(x,15,replace = T)
 [1]  7  4  7  2  7  7  9  7  6  2  3  4  6 6 10
```

(**i**)

```
> d1 <- data.frame(v1,v3)
> d1
  v1     v3
1  2    low
2  3 medium
3  5   high
> rm(v1,v3)
> v1 # v1 now exists only internal to d1
Error: Object "v1" not found
> search()
 [1] "C:\\Program Files\\sp2000\\users\\stauffer\\_Data"
 [2] "C:\\Program Files\\sp2000\\splus\\_Functio"
 [3] "C:\\Program Files\\sp2000\\stat\\_Functio"
 [4] "C:\\Program Files\\sp2000\\s\\_Functio"
 …
> attach(d1)
> search()
 [1] "C:\\Program Files\\sp2000\\users\\stauffer\\_Data"
 [2] "d1"
 [3] "C:\\Program Files\\sp2000\\splus\\_Functio"
 [4] "C:\\Program Files\\sp2000\\stat\\_Functio"
 …
> v1 # v1 now exists at level 2 in the directory
[1] 2 3 5
> detach(2)
> v1 # again, v1 now exists only internal to d1
Error: Object "v1" not found
```

Figure 1.3. *Continued.*

(**j**)

```
function(x)
{
# function: add - adds the values in vector x
# author: Jill Analyst
# date:   January 1, 2007
sum <- 0.0
for (i in 1:length(x))
   {sum <- sum + x[i]}
return(sum)
}
```

(**k**)

```
> v1
[1] 2 3 5
> add(v1)
[1] 10
```

(**l**)

```
> x <- runif(20,2,8)
> y <- 10+1.5*x+rnorm(20,0,1)
> output <- lm(y~x)
> summary(output)
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 10.1862  0.8926    11.4113  0.0000
          x  1.4600  0.1655     8.8207  0.0000
Residual standard error: 1.179 on 18 degrees of freedom
Multiple R-Squared: 0.8121
F-statistic: 77.8 on 1 and 18 degrees of freedom, the p-value is
     5.937e-008
> plot(x,y) # See Fig. 1.4.
> abline(10.1862,1.4600)
```

S-Plus and R (hereafter we omit reference to the <ENTER> key at the end of each input command where this is clear from the context) (Fig. 1.3b). All of these data object reponses are vectors of values, starting with the first entry indexed by `[1]` at the beginning of the output line. S-Plus and R are case-sensitive. Use the `Help` menu `Search S-Plus Help` in S-Plus or the `Help` menu `R functions (text) ...` option in R to learn more about the commands and their syntax.

The `<-` operation assigns values to objects, and the `rm` command removes or deletes objects in both S-Plus and R (Fig. 1.3c). In S-Plus, the underscore _ can be substituted for `<-` to indicate assignment. The list of objects currently available at the top-level directory of S-Plus and R can be viewed by using the `objects()` command.

1.6.3 Data Structures

In S-Plus and R, datasets are organized into simple data structures of type **numeric** for quantitative or numerical values, **factor** for qualitative or categorical values, **character** for character strings, and **logical** for objects with logical values true `T` or false `F`. These simple atomic types can be combined into complex data structures such as one-dimensional **vectors** of simple values of the same type, two-dimensional **matrices** of columns of vectors of the same type and length, two-dimensional **data frames** with columns of vectors of possibly different types but the same size, and **lists** of simple and complex types of varying size. Vectors are matrices, matrices are data frames, and data frames are lists, but the converse is rarely true. Let's combine simple values into a vector with the concatenate `c` command, vectors into matrices with the `matrix` command, column bind `cbind` command, or row bind `rbind` command, vectors into data frames with the `data.frame` command, and data structures into lists with the `list` command (Fig. 1.3d). Notice how the row entries in the matrix `m1` are indicated by `[1,]`, `[2,]`, and `[3,]`; the column vector entries in the data frame `d1` are indicated by `v1` and `v3`; and the entries in the list `list1` are indicated by `[[1]]` and `[[2]]`. The type of values in data structures can be tested with the `is.numeric`, `is.character`, `is.logical`, `is.vector`, `is.matrix`, `is.data.frame`, or `is.list` commands (Fig. 1.3e). Data structures can sometimes be coerced into other types of more complex objects and values with `as.matrix` to convert a vector into a matrix, `as.data.frame` to convert a matrix into a data frame, `as.list` to convert a data frame into a list, and `as.character` to convert a numeric vector into a character vector (Fig. 1.3f).

1.6.4 Random Numbers

Probability density, cumulative probability, quantile, and random values can be generated in S-Plus and R by using the `d`, `p`, `q`, and `r` prefixes with common probability distributions such as the normal, uniform, gamma, exponential, *t*, *F*, chi-square (χ^2), lognormal, Poisson, negative binomial, and binomial distributions using `norm`, `unif`, `gamma`, `exp`, `t`, `f`, `chisq`, `lnorm`, `pois`, `nbinom`, and `binom` suffixes and their specified parameters (Fig. 1.3g). For example, in the figure, *y* is a vector with 10 numeric values randomly sampled from the normal distribution $N(5, 1)$ with mean $\mu = 5$ and standard deviation $\sigma = 1$. The cumulative probability at 4, quantile of 0.1586553, and density value at 5 for $N(5, 1)$ are also calculated.

The `sample` command can be used to generate random data from a vector, without or with replacement (Fig. 1.3h). The default of this command is to sample without replacement.

1.6.5 Graphs

Graphs can be obtained for numeric values using the `dotplot` (`stripchart` in R) and `hist` commands on vectors, `plot` command on pairs of vectors, and `pairs` command on matrices or data frames of columns. You can also create

an $n \times m$ trellis of graphs with n rows and m columns using the `par(mfrow=c(n,m))` command.

1.6.6 Importing and Exporting Files

Data files can be imported or exported as objects using the `Import Data` and `Export Data` options in the `File` menu in S-Plus. Data files are imported as data frames unless specified otherwise. In R, data can be imported and exported using `copy and paste` commands in the data file and the `Data editor` option in the `Edit` menu. Alternatively, in R, text files of data with the `txt` extension `data.txt` from the defaulted R-2.2.1 folder can be imported using the `read.table("data.txt")` command.

1.6.7 Saving and Restoring Objects

You can save and restore your directory of objects or individual objects in S-Plus by using `data.dump(objects(), "directory and filename")` or `data.dump(c("object1","object2",...), "directory and filename")` commands and `data.restore("directory and filename")` commands. Alternatively, in S-Plus, you can use the `Workspace Save` and `Workspace Open` options in the `File` menu. In R, you can save and restore your directory of objects by using the `Save Workspace` and `Load Workspace` options in the `File` menu.

1.6.8 Directory Structure

The S-Plus and R directory structures are hierarchical in structure and can be viewed using the `search()` command (Fig. 1.3i). The directory of an object can be determined using the `find (object)` command. An object such as a data frame or list can be opened and closed, with internal objects such as vectors made available at a specified directory level (defaulted to level 2), by using the `attach(object, level)` and `detach(level)` commands. The # symbol in a line of code indicates to S-Plus or R that the remainder of the command is a comment.

1.6.9 Functions and Control Structures

Function subprograms can be created as objects using the `fix(name of object)` command. If a mistake is made in creating and editing the subprogram object, an error message will be issued. Use the `fix()` command to return to editing to correct the mistake in the object before signing off from S-Plus or R. Otherwise the function object will not be saved. The standard programming control structures are available in S-Plus and R: (1) sequential; (2) conditional, with `if (test) then {} else {}` syntax; (3) repetition, with `for (comparison or name in values) expr` or `for (comparison or name in values) expr else`