1
Introduction

1.1 Objectives of the Book

This book is about making valid inferences from scientific data when a meaningful analysis depends on a model of the information in the data. Our general objective is to provide scientists, including statisticians, with a readable text giving practical advice for the analysis of empirical data under an information-theoretic paradigm. We first assume that an exciting scientific question has been carefully posed and relevant data have been collected, following a sound experimental design or probabilistic sampling program. Alternative hypotheses, and models to represent them, should be carefully considered in the design stage of the investigation. Often, little can be salvaged if data collection has been seriously flawed or if the question was poorly posed (Hand 1994). We realize, of course, that these issues are never as ideal as one would like. However, proper attention must be placed on the collection of data (Chatfield 1991, 1995a Anderson 2001). We stress inferences concerning the structure and function of biological systems, relevant parameters, valid measures of precision, and formal prediction.

There are many studies where we seek an understanding of relationships, especially causal ones. There are many studies to understand our world; models are important because of the parameters in them and relationships expressed between and among variables. These parameters have relevant, useful interpretations, even when they relate to quantities that are not directly observable (e.g., survival probabilities, animal density in an area, gene frequencies, and interaction terms). Science would be very limited without such unobservables

as constructs in models. We make statistical inferences from the data, to a real or conceptual population or process, based on models involving such parameters. Observables and prediction are often critical, but science is broader than these issues.

The first objective of this book is to outline a consistent *strategy* for issues surrounding the analysis of empirical data. Induction is used to make statistical inference about a defined population or process, given an empirical sample or experimental data set. "Data analysis" leading to valid inference is the integrated process of careful a priori model formulation, model selection, parameter estimation, and measurement of precision (including a variance component due to model selection uncertainty). We do not believe that model selection should be treated as an activity that precedes the analysis; rather, model selection is a critical and integral aspect of scientific data analysis that leads to valid inference.

A philosophy of thoughtful, science-based, a priori modeling is advocated. Often, one first develops a global model (or set of models) and then derives several other plausible candidate (sub)models postulated to represent good approximations to information in the data at hand. This forms the *set of candidate models*. Science and biology play a lead role in this a priori model building and careful consideration of the problem. A simple example of models to represent alternative scientific hypotheses might be helpful at this early point. Consider the importance of an interaction between age (a) and winter severity (w) in a particular animal population. A model including such an interaction would have the main effects plus the interaction; $a + w + a * w$, while the model $a + w$ lacks the interaction term. Information-theoretic methods allow several lines of quantitative evidence concerning the importance of this hypothesized interaction.

The modeling and careful thinking about the problem are critical elements that have often received relatively little attention in statistics classes (especially for nonmajors), partly because such classes rarely consider an overall strategy or philosophy of data analysis. A proper a priori model-building strategy tends to avoid "data dredging," which leads to overfitted models, that is, to the "discovery" of effects that are actually spurious (Anderson 2001a). Instead, there has often been a rush to "get to the data analysis" and begin to rummage through the data and compute various estimates of interest or conduct null hypothesis tests. We realize that these other philosophies may have their place, especially in more exploratory investigations.

The second objective is to explain and illustrate methods developed recently at the interface of information theory and mathematical statistics for selection of an estimated "best approximating model" from the a priori set of candidate models. In particular, we review and explain the use of Akaike's information criterion (AIC) in the selection of a model (or small set of good models) for statistical inference. AIC provides a simple, effective, and objective means for the selection of an estimated "best approximating model" for data analysis and inference. Model selection includes "variable selection" as frequently

practiced in regression analysis. Model selection based on information theory is a relatively new paradigm in the biological and statistical sciences and is quite different from the usual methods based on null hypothesis testing. Model selection based on information theory is not the only reasonable approach, but it is what we are focusing on here because of its philosophical and computational advantages.

The practical use of information criteria, such as Akaike's, for model selection is relatively recent (the major exception being in time series analysis, where AIC has been used routinely for the past two decades). The marriage of information theory and mathematical statistics started with Kullback's (1959) book. Akaike considered AIC to be an extension of R. A. Fisher's likelihood theory. These are all complex issues, and the literature is often highly technical and scattered widely throughout books and research journals. Here we attempt to bring this relatively new material into a readable text for people in (primarily) the biological and statistical sciences. We provide a series of examples, many of which are biological, to illustrate various aspects of the theory and application.

In contrast, hypothesis testing as a means of selecting a model has had a much longer exposure in science. Many seem to feel more comfortable with the hypothesis testing paradigm in model selection, and some even consider the results of a test as *the* standard by which other approaches should be judged (we believe that they are wrong to do so). Bayesian methods in model selection and inference have been the focus of much recent research. However, the technical level of this material often makes these approaches unavailable to many in the biological sciences. A variety of cross-validation and bootstrap-based methods have been proposed for model selection, and these, too, seem like very reasonable approaches. The computational demands of many of the Bayesian and cross-validation methods for model selection are often quite high (often 1–3 orders of magnitude higher than information-theoretic approaches), especially if there are more than a dozen or so high-dimensional candidate models.

The theory presented here allows estimates of "model selection uncertainty," inference problems that arise in using the same data for both model selection and the associated parameter estimation and inference. If model selection uncertainty is ignored, precision is often overestimated, achieved confidence interval coverage is below the nominal level, and predictions are less accurate than expected. Another problem is the inclusion of spurious variables, or factors, with no assessment of the reliability of their selection. Some general methods for dealing with model- and variable-selection uncertainty are suggested and examples provided. Incorporating model selection uncertainty into estimators of precision is an active area of research, and we expect to see additional approaches developed in the coming years.

The third objective is to present a number of approaches to making formal inference from more than one model in the set. That is, rather than making inferences from only the model estimated to be the best, robust inferences can

be made from several, even all, models being considered. These procedures are termed *multimodel inference* (MMI). Model averaging has been an active research area for Bayesians for the past several years (Hoeting et al. 1999). Model averaging can be easily done under an information-theoretic approach. Model averaging has several practical and theoretical advantages, particularly in prediction or in cases where a parameter of interest occurs in all the models. Confidence sets on models is another useful approach, particularly when models in the set represent a logical ordering (e.g., a set of models representing chronic treatment effects over 1, 2, . . . , t time periods). Finally, the relative importance of explantory variables in a general regression setting can be easily assessed by summing certain quantities across models. MMI is also potentially useful in certain conflict resolution issues (Anderson et al. 2001c).

Current practice often would judge a variable as important or unimportant, based on whether that variable was in or out of the selected model (e.g., stepwise regression, based on hypothesis testing). Such procedures provide a misleading dichotomy (see Breiman 2001) and are not in the spirit of a weight of evidence. MMI allows us to discard simplistic dichotomies and focus on quantitatively ranking models and variables as to their relative value and importance.

Modeling is an art as well as a science and is directed toward finding a good approximating model of the information in empirical data as the basis for statistical inference from those data. In particular, the number of parameters estimated from data should be substantially less than the sample size, or inference is likely to remain somewhat preliminary (e.g., Miller (1990: x)) mentions a regression problem with 757 variables and a sample size of 42 (it is absurd to think that valid inference is likely to come from the analysis of these data). In cases where there are relatively few data per estimated parameter, a small-sample version of AIC is available (termed AIC_c) and should be used routinely rather than AIC. There are cases where quasi-likelihood methods are appropriate when count data are overdispersed; this theory leads to modified criteria such as QAIC and QAIC_c, and these extensions are covered in the following material.

Simple models with only 1-2 parameters are not the central focus of this book; rather, we focus on models of more complex systems. Parameter estimation has been firmly considered to be an optimization problem for many decades, and AIC formulates the problem of model selection as an optimization problem across a set of candidate models. Minimizing AIC is a simple operation with results that are easy to interpret. Models can be clearly ranked and scaled, allowing full consideration of other good models, in addition to the estimated "best approximating model." Evidence ratios allow a formal strength of evidence for alternative hypotheses. Competing models, those with AIC values close to the minimum, are also useful in the estimation of model selection uncertainty. Inference should often be based on more than a single model, unless the data clearly support only a single model fit to the data. Thus, some approaches are provided to allow inference from several or all of the models, including model averaging.

This is primarily an applied book. A person with a good background in mathematics and theoretical statistics would benefit from studying Chapter 7. McQuarrie and Tsai (1998) present both theoretical and applied aspects of model selection in regression and time series analysis, including extensive results of large-scale Monte Carlo simulation studies.

1.2 Background Material

Data and stochastic models of data are used in the empirical sciences to make inferences concerning both processes and parameters of interest (see Box et al. 1981, Lunneborg 1994, and Shenk and Franklin 2001 for a review of principles). Statistical scientists have worked with researchers in the biological sciences for many years to improve methods and understanding of biological processes. This book provides practical, omnibus methods to achieve valid inference from models that are good approximations to biological processes and data. We focus on statistical evidence and try to avoid arbitrary dichotomies such as "significant or not significant." A broad definition of data is employed here. A single, simple data set might be the subject of analysis, but more often, data collected from several field sites or laboratories are the subject of a more comprehensive analysis. The data might commonly be extensive and partitioned by age, sex, species, treatment group, or within several habitat types or geographic areas. In linear and nonlinear regression models there may be many explanatory variables. There are often factors (variables) with small, moderate, and large effects in these information-rich data sets (the concept of tapering effect sizes). Parameters in the model represent the effects of these factors. We focus on modeling philosophy, model selection, estimation of model parameters, and valid measures of precision under the relatively new paradigm of information-theoretic methods. Valid inference rests upon these four issues, in addition to the critical considerations relating to problem formulation, study design, and protocol for data collection.

1.2.1 Inference from Data, Given a Model

R. A. Fisher (1922) discussed three aspects of the general problem of valid inference: (1) model specification, (2) estimation of model parameters, and (3) estimation of precision. Here, we prefer to partition model specification into two components: formulation of a set of candidate models and selection of a model (or small number of models) to be used in making inferences. For much of the twentieth century, methods have been available to objectively and efficiently estimate model parameters and their precision (i.e., the sampling covariance matrix). Fisher's *likelihood theory* has been the primary omnibus approach to these issues, but it *assumes* that the model structure is known (and correct, i.e., a true model) and that only the parameters in that structural

model are to be estimated. Simple examples include a linear model such as $y = \alpha + \beta x + \epsilon$ where the residuals (ϵ) are assumed to be normally distributed, or a log-linear model for the analysis of count data displayed in a contingency table. The parameters in these models can be estimated using *maximum likelihood* (ML) methods. That is, if one assumes or somehow chooses a particular model, methods exist that are objective and asymptotically optimal for estimating model parameters and the sampling covariance structure, conditional on that model. A more challenging example might be to assume that data are appropriately modeled by a 3-parameter gamma distribution; one can routinely use the method of maximum likelihood to estimate these model parameters and the model-based 3×3 sampling covariance matrix. Given an appropriate model, and if the sample size is "large," then maximum likelihood provides estimators of parameters that are consistent (i.e., asymptotically unbiased with variance tending to zero), fully efficient (i.e., minimum variance among consistent estimators), and normally distributed. With small samples, but still assuming an appropriate model, ML estimators often have small-sample bias, where bias $\equiv \mathrm{E}(\hat{\theta}) - \theta$. Such bias is usually a trivial consideration, as it is often substantially less than the se($\hat{\theta}$), and bias-adjusted estimators can often be found if this is deemed necessary. The sampling distributions of ML estimators are often skewed with small samples, but profile likelihood intervals or log-based intervals or bootstrap procedures can be used to achieve asymmetric confidence intervals with good coverage properties. **In general, the maximum likelihood method provides an objective, omnibus theory for estimation of model parameters and the sampling covariance matrix, *given an appropriate model*.**

1.2.2 Likelihood and Least Squares Theory

Biologists have typically been exposed to least squares (LS) theory in their classes in applied statistics. LS methods for linear models are relatively simple to compute, and therefore they enjoyed an early history of application (Weisburg 1985). In contrast, Fisher's likelihood methods often require iterative numerical methods and were thus not popular prior to the widespread availability of personal computers and the development of easy-to-use software. LS theory has many similarities with likelihood theory, and it yields identical estimators of the structural parameters (but not σ^2) for linear and nonlinear models when the residuals are assumed to be independent and normally distributed. It is now easy to allow alternative error structures (i.e., nonnormal residuals such as Poisson, gamma or log-normal) for regression and other similar problems in either a likelihood or quasi-likelihood framework (e.g., McCullagh and Nelder 1989, Heyde 1997), but more difficult in an LS framework.

The concepts underlying both estimation methods are relatively simple to understand (Silvey 1975). Consider the simple linear regression, where a response variable (y) is modeled as a linear function of an explanatory variable

(x) as $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$. The ϵ_i are error terms (residuals) which are often modeled as independent normal random variables with mean 0 and constant variance σ^2. Under LS the estimates of β_0 and β_1 are those that minimize $\sum(\epsilon_i)^2-$ hence the name *least squares*. The parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the average squared error terms (ϵ_i) and define a regression line that is the "best fit." Hundreds of statistics books cover the theory and application for least squares estimation in linear and nonlinear models, particularly when the ϵ_i are assumed to be independent, normally distributed random variables.

Likelihood methods are much more general, far less taught in applied statistics courses, and slightly more difficult to understand at first. The material in much of this book relies on an understanding of likelihood theory, so some brief introduction is given here. While likelihood theory is a paradigm underlying both frequentist and Bayesian statistics, there are no more than a handful of applied books solely on this important subject (good examples include McCullagh and Nelder 1989, Edwards 1992, Azzalini 1996, Morgan 2000, and Severini 2000).

The theory underlying likelihood begins with a probability model, given the parameters (θ). Specifically, model g describes the probability distribution of the data, given the model parameters and a specific model form; denoted by $g(x|\theta, \mathit{model})$. A simple example is the binomial probability function where θ is the probability of a "success"; let this be the parameter $p = 0.4$. The data could be the observation of $y = 15$ successes out of $n = 40$ independent trials. Then, the discrete probability of getting 15 successes out of 40 trials, given the parameter $(p \equiv 0.4)$ and the binomial model, is

$$g(y, n|p, \mathit{binomial}) = \binom{n}{y} p^y (1-p)^{n-y},$$

$$g(15, 40|p = 0.4, \mathit{binomial}) = \frac{40!}{15!25!}(0.4)^{15}(1-0.4)^{25} = 0.123.$$

The key point is that for this calculation, the model (here a binomial model) and its parameters (here $p = 0.4$) are known in advance (i.e., they are *given*). In very simple problems such as this, an excellent model is available and can be considered given (such is rarely the case in the real world, where one is not sure what model might be used). Then one observes the data ($y = 15$ and $n = 40$) and can compute the probability of the data, given the model and its parameters.

In much of science, neither the model parameters nor the model is known. However, data can be collected in a way that allows the parameters to be estimated if a good model can be found or assumed. The likelihood function is the basis for such parameter estimation and is a function of the parameter p, given the data and the binomial model:

$$\mathcal{L}(p|y, n, \mathit{binomial}) = \binom{n}{y} p^y (1-p)^{n-y}$$

or

$$\mathcal{L}(p|15, 40, \textit{binomial}) = \frac{40!}{15!25!}(p)^{15}(1-p)^{25}.$$

Clearly, the likelihood is a function of (only) the unknown parameter (p in this example); everything else is known or assumed. The probability model and the associated likelihood function differ only in terms of what is known or given. In the probability model, the parameters, the model, and the sample size are known, and interest lies in the probability of observing a particular event (the data, y given n in this simple example). In the likelihood function, the data are given (observed) and the model is assumed (but given), and interest lies in estimating the unknown parameters; thus, the likelihood is a function of only the parameters. The probability model of the data and the likelihood function of the parameters are closely related; they merely reverse the roles of the data and the parameters, given a model. The binomial coefficient $\binom{n}{y}$ does not contain the unknown parameter p and is often omitted (it does not contain any information about the unknown parameters and is often difficult to compute if $n > 50$).

The notation for the likelihood function is very helpful in its understanding; consider the general expression $\mathcal{L}(\theta|\textit{data}, \textit{model})$. If we follow the usual convention of letting x represent the empirical data and g a *given* approximating model, then $\mathcal{L}(\theta|x, g)$ is read as "the likelihood of a particular numerical value of the unknown parameter θ (θ is usually a vector), given the data x and a particular model g."

A well-known example will help illustrate the concept. Consider flipping n pennies and observing y "heads." Assuming that the flips are independent and that each penny has an equal probability of a head, the binomial model is an obvious model choice in this simple setting. The likelihood function is $\mathcal{L}(p|y, n,$ binomial), where p is the (unknown) probability of a head. Thus, given the data (y and n) and the binomial model, one can compute the *likelihood* that p is 0.15 or 0.73 or any other value between 0 and 1. The likelihood (a relative, not absolute, value) is a function of the unknown parameter p. Given this formalism, one might compute the likelihood of many values of the unknown parameter p and pick the most likely one as the *best* estimate of p, *given* the data and the model. It seems compelling to pick the value of p that is "most likely." This is Fisher's concept of *maximum likelihood estimation*; he published this when he was 22 years old as a third-year undergraduate at Cambridge University! He reasoned that the best estimate of an unknown parameter (given data and a model) was that which was the most likely; thus the name *maximum likelihood*, ML. The ML estimate (MLE) for the binomial model happens to have a closed-form expression that is well known: $\hat{p} = y/n = 7/11 = 0.6363$. That is, the numerical value of y/n exactly maximizes the likelihood function. In most real-world cases a simple, closed form estimator either does not exist or cannot be found without substantial difficulty.

Likelihood theory includes asymptotically optimal methods for estimation of unknown parameters and their variance–covariance matrix, derivation of hypothesis tests, the basis for profile likelihood intervals, and other important quantities (such as model selection criteria). More generally, likelihood theory includes the broad concept of *support* (Edwards 1992). Likelihood is also the essential basis for Bayesian approaches to statistical inference. In fact, likelihood is the backbone of statistical theory, whereas least squares can be viewed as a limited special case and, while very useful in several important applications, is not foundational in modern statistics.

For many purposes the natural logarithm of the likelihood function is essential; written as $\log(\mathcal{L}(\theta|data, model))$, or $\log(\mathcal{L}(\theta|x, model))$, or if the context is clear, just $\log(\mathcal{L}(\theta))$ or even just $\log(\mathcal{L})$. Often, one sees notation such as $\log(\mathcal{L}(\theta|x))$, without it being clear that a particular model is assumed. An advanced feature of $\log(\mathcal{L})$ is that it, by itself, is a type of *information* concerning θ and the model (Edwards 1992:22–23). The log-likelihood for the binomial model where 11 pennies are flipped and 7 heads are observed is

$$\begin{aligned}\log(\mathcal{L}(p|y, n, binomial)) &= \log\binom{n}{y} + y \cdot \log(p) + (n - y) \cdot \log(1 - p),\\ &= \log\binom{11}{7} + 7 \cdot \log(p) + (11 - 7) \cdot \log(1 - p)\\ &= 5.79909 + 7 \cdot \log(p) + (4) \cdot \log(1 - p).\end{aligned}$$

A property of logarithms for values between 0 and 1 is that they lie in the negative quadrant; thus, values of discrete log-likelihood functions are negative (unless some additive constants have been omitted). Figure 1.1 shows a plot of the likelihood (a) and log-likelihood (b) functions where 11 pennies were flipped, 7 heads were observed, and the binomial model was assumed. The value of $p = 0.636$ maximizes both the likelihood and the log-likelihood function; this value is denoted by $\hat{p}$ and is the maximum likelihood estimate (MLE). Relatively little information is contained in such a small sample size ($n = 11$) and this is reflected in the broad shape of the plots. Had the sample size been 5 times larger, with $n = 55$ and 35 heads observed, the likelihood and log-likelihood functions would be more peaked (Figure 1.1c and d). In fact, the sampling variance is derived from the shape of the log-likelihood function around its maximum point. In the usual case where θ is a vector, a variance–covariance matrix can be estimated based on partial derivatives of the log-likelihood function. These procedures will not be developed here.

The value of the log-likelihood function at its maximum point is a very important quantity, and it is this point that defines the *maximum likelihood estimate*. In the example with 11 flips and 7 heads, the value of the maximized log-likelihood is -1.411 (Figure 1.1b). This result is computed by taking the

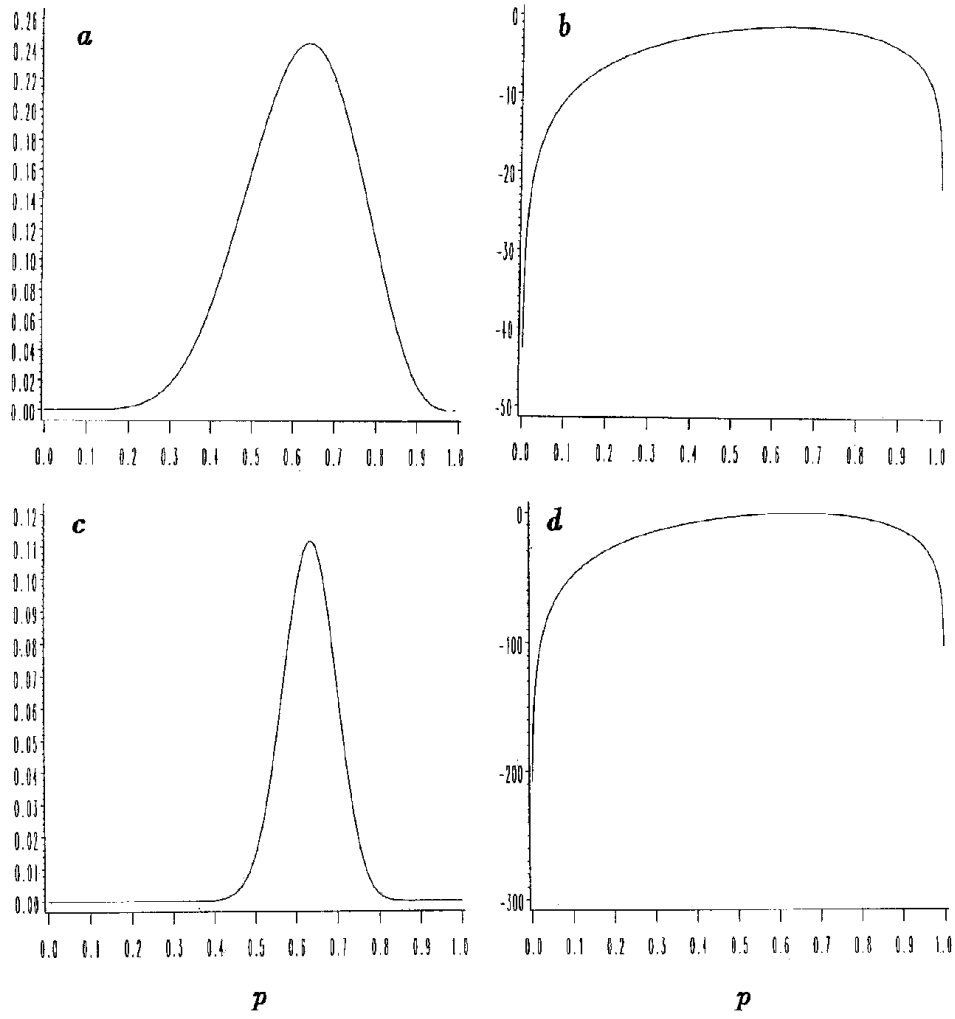

FIGURE 1.1. Plots of the binomial likelihood (a) and log-likelihood (b) function, given $n = 11$ penny flips and the observation that $y = 7$ of these were heads. Also shown are plots of the binomial likelihood (c) and log-likelihood (d) function, given a sample size 5 times larger; $n = 55$ penny flips and the observation that $y = 35$ of these were heads. Note the differing scales on the Y axis.

log-likelihood function

$$\log(\mathcal{L}(p|y, n, \textit{binomial})) = \log\binom{n}{y} + y \cdot \log(p) + (n - y) \cdot \log(1 - p)$$

and substituting the MLE ($\hat{p} = 0.6363$) and the data (y and n),

$$-1.411 = 5.79909 + 7 \cdot \log(0.6363) + (4) \cdot \log(1 - 0.6363).$$

Thus, when one sees reference to a maximized $\log(\mathcal{L}(\theta))$ this merely represents a numerical value (e.g., -1.411).

Many do not realize that the common procedure for setting a 95% confidence interval (i.e., $\hat{\theta} \pm 1.96 \cdot \widehat{\text{se}}(\hat{\theta})$) is merely an approximation. The estimator $\hat{\theta}$ is

only asymptotically normal, and if the sample size is too small, the sampling distribution will often be nonnormal and the approximation will be poor (i.e., achieved confidence interval coverage can be much less than the nominal value, say, 95%). For example, if the binomial parameter is near 0 or 1, the distribution of the estimator $\hat{\theta}$ will be nonnormal (asymmetric) unless the sample size is very large. In general, rather than use the simple approximation, one can set a 95% interval using the log-likelihood function; this procedure, in general, is called a *profile likelihood interval*. This is not a simple procedure; thus the approximation has seen heavy use in applied data analysis. We cannot provide the full theory for profile likelihood intervals here, but will give an example for the binomial case where $n = 11$, $y = 7$, $\hat{p} = 0.6363$, and the maximized log-likelihood value is -1.411. Here, we start with 3.84, which is the 0.05 point of the chi-squared distribution with 1 degree of freedom. One-half of this value is 1.92, and this value is subtracted from the maximum point of the log-likelihood function: $-1.411 - 1.92 = -3.331$. Now, numerically, one must find the 2 values of p that are associated with the values of the log-likelihood function at -3.331. These 2 values are the endpoints of an exact 95% likelihood confidence interval. In this example, the 95% likelihood interval is (0.346, 0.870).

Biologists familiar with LS but lacking insight into likelihood methods might benefit from an example. Consider a multiple linear regression model where a dependent variable y is hypothesized to be a function of r explanatory (predictor) variables x_j ($j = 1, 2, \ldots, r$). Here the residuals ϵ_i of the n observations are assumed to be independent, normally distributed with a constant variance σ^2, and the model structure is expressed as

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \epsilon_i, \qquad i = 1, \ldots, n.$$

Hence

$$E(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r, \qquad i = 1, \ldots, n,$$

and $E(y_i)$ is a linear function of $r + 1$ parameters. The conceptual residuals,

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r) = y_i - E(y_i),$$

have the joint probability distribution $g(\underline{\epsilon}|\underline{\theta})$, where $\underline{\theta}$ is a vector of $K = r + 2$ parameters ($\beta_0, \beta_1, \ldots, \beta_r$, and σ). Here, corresponding to observation i one has the model

$$g(\epsilon_i|\underline{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left[\frac{\epsilon_i}{\sigma}\right]^2}.$$

The likelihood is simply the product of these over the n observations, interpreted as a function of the unknown parameters, given the data, the linear model structure, and the normality assumption:

$$\mathcal{L}(\underline{\theta}|\underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left[\frac{\epsilon_i}{\sigma}\right]^2} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{\epsilon_i}{\sigma}\right]^2}.$$

Here we use "x" in $\mathcal{L}(\underline{\theta}|x)$ to denote the full data. When the ϵ_i are normally distributed with constant variance σ^2, the maximum likelihood estimator (MLE) of $\underline{\beta}$ is identical to the usual LS regression estimators (however, the estimator of σ^2 differs slightly). This formalism shows, *given the model*, the link between the data, the model, and the parameters to be objectively estimated, using either LS or ML.

In all fitted linear models the residual sum of squares (RSS) is

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2,$$

where

$$\begin{aligned} \hat{\epsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_r x_r), \\ &= y_i - \hat{\text{E}}(y_i) \end{aligned}$$

The ML estimator is $\hat{\sigma}^2 = \text{RSS}/n$, while the estimator universally used in the LS case is $\hat{\sigma}^2 = \text{RSS}/(n-(r+1))$. This shows that ML and LS estimators of σ^2 differ by a factor of $n/(n-(r+1))$; often a trivial difference unless the sample size is small. The maximized likelihood is

$$\mathcal{L}(\hat{\underline{\theta}}|x) = \left[\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right]^n e^{-\frac{1}{2}n},$$

or

$$\log(\mathcal{L}(\hat{\underline{\theta}})) = -\frac{1}{2}n\log(\hat{\sigma}^2) - \frac{n}{2}\log(2\pi) - \frac{n}{2}.$$

The additive constants can often be discarded from the log-likelihood because they are constants that do not influence likelihood-based inference. Thus for all standard linear models, we can take

$$\log(\mathcal{L}(\hat{\underline{\theta}})) \approx -\frac{1}{2}n\log(\hat{\sigma})^2.$$

This result is important in model selection theory because it allows a simple mapping from LS analysis results (e.g., the RSS or the MLE of σ^2) into the maximized value of the log-likelihood function for comparisons over such linear models with normal residuals. Note that the log-likelihood is defined up to an arbitrary additive constant in this usual case. If the model set includes linear and nonlinear models or if the residual distributions differ (e.g., normal, gamma, and log-normal), then all the terms in the log-likelihood must be retained, without omitting any constants. Most uses of the log-likelihood are relative to its maximum, or to other likelihoods at their maxima, or to the curvature of the log-likelihood function at the maximum.

The number of parameters $K = r+2$ in these linear models must include the intercept (say, β_0), the r regression coefficients ($\beta_1, \ldots, \beta_r$), and the residual variance (σ^2). Often, one (erroneously) considers only the number of parameters being estimated as the intercept and the slope parameters (ignoring σ^2);

Sir Ronald Aylmer Fisher was born in 1890 in East Finchley, London, and died in Adelaide, Australia, in 1962. This photo was taken when he was approximately 66 years of age. Fisher was one of the foremost scientists of his time, making incredible contributions in theoretical and applied statistics and genetics. Details of his life and many scientific accomplishments are found in Box (1978). He published 7 books (one of these had 14 editions and was printed in 7 languages) and nearly 300 journal papers. Most relevant to the subject of this book is Fisher's likelihood theory and parameter estimation using his method of maximum likelihood.

however, in the context of model selection, the number of parameters must include σ^2 and thus $K = r + 2$. If the method of LS is used to obtain parameter estimators, one must use the regression-based estimate of σ^2 times $(n - (r + 1))/n = (n - K + 1)/n$ to obtain the ML estimator of σ^2. In LS estimation, we minimize RSS $= n\hat{\sigma}^2$, which for all parameters other than σ^2 itself is equivalent to maximizing $-\frac{1}{2} \cdot n \log(\hat{\sigma}^2)$.

There is a close relationship between LS and ML methods for linear and nonlinear models, where the ϵ_i are assumed to be normally distributed. For example, the LS estimates of the structural model parameters (but not σ^2) are equivalent to the MLEs. Likelihood (and related Bayesian) methods allow easy extensions to the many other classes of models and, with the exploding power of computing equipment, likelihood methods are finding increasing use by both statisticians and researchers in other scientific disciplines (see Garthwaite et al. 1995 for background).

1.2.3 The Critical Issue: "What Is the Best Model to Use?"

While hundreds of books and countless journal papers deal with estimation of model parameters and their associated precision, relatively little has appeared

concerning model specification (what set of candidate models to consider) and model selection (what model(s) to use for inference) (see Peirce 1955). In fact, Fisher believed at one time that model specification was outside the field of mathematical statistics, and this attitude prevailed within the statistical community until at least the early 1970s. "*What is the best model to use*?" is *the* critical question in making valid inference from data in the biological sciences.

The likelihood function $\mathcal{L}(\theta|x, \textit{model})$ makes it clear that for inference about θ, data and the model are taken as *given*. Before one can compute the likelihood that $\theta = 5.3$, one must have data and a particular statistical model. While an investigator will have empirical data for analysis, it is unusual that the model is known or given. Rather, a number of alternative model forms must be somehow considered as well as the specific explanatory variables to be used in modeling a response variable. This issue includes the *variable selection problem* in multiple regression analysis. If one has data and a model, LS or ML theory can be used to estimate the unknown parameters (θ) and other quantities useful in making statistical inferences. However, which model is the best to use for making inferences? What is the basis for saying a model is "best"?

Model selection relates to fitted models: given the data and the form of the model, then the MLEs of the model parameters have been found ("fitted"). Inference relates to theoretical models. It is necessary to consider four cases;

(1) models as structure only (θ value irrelevant),
(2) models as structure, plus specific θ_o (this is the theoretical best value),
(3) models as structure, plus MLE $\hat{\theta}$, fitted to data,
(4) models as structure by fitting, downplaying θ.

If a poor or inappropriate model (3, above) is used, then inference based on the data and this model will often be poor. Thus, it is clearly important to select (i.e., infer) an appropriate model (1, above) for the analysis of a specific data set; however, this is not the same as trying to find the "true model." Model selection methods with a deep level of theoretical support are required and, particularly, methods that are easy to use and widely applicable in practice. Part of "applicability" means that the methods have good operating characteristics for realistic sample sizes. As Potscher (1991) noted, asymptotic properties are of little value unless they hold for realized sample sizes.

A simple example will motivate some of the concepts presented. Flather (1992 and 1996) studied patterns of avian species-accumulation rates among forested landscapes in the eastern United States using index data from the Breeding Bird Survey (Bystrak 1981). He derived an a priori set of 9 candidate models from two sources: (1) the literature on species area curves (most often the power or exponential models were suggested) and (2) a broader search of the literature for functions that increased monotonically to an asymptote (Table 1.1). Which model should be used for the analysis of these ecological data? Clearly, none of these 9 models are likely to be the "truth" that generated

TABLE 1.1. Summary of a priori models of avian species-accumulation curves from Breeding Bird Survey index data for Indiana and Ohio (from Flather 1992:51 and 1996). The response variable (y) is the number of accumulated species, and the explanatory variable (x) is the accumulated number of samples. Nine models and their number of parameters are shown to motivate the question, "Which fitted model should be used for making inference from these data?"

Model structure	Number of parameters (K)[a]
$\mathrm{E}(y) = ax^b$	3
$\mathrm{E}(y) = a + b\log(x)$	3
$\mathrm{E}(y) = a(x/(b+x))$	3
$\mathrm{E}(y) = a(1 - e^{-bx})$	3
$\mathrm{E}(y) = a - bc^x$	4
$\mathrm{E}(y) = (a + bx)/(1 + cx)$	4
$\mathrm{E}(y) = a(1 - e^{-bx})^c$	4
$\mathrm{E}(y) = a\left(1 - [1 + (x/c)^d]^{-b}\right)$	5
$\mathrm{E}(y) = a[1 - e^{-(b(x-c))^d}]$	5

[a] There are $K-1$ structural parameters and one residual variance parameter, σ^2. Assumed: $y = \mathrm{E}(y)+\epsilon$, $\mathrm{E}(\epsilon) = 0$, $\mathrm{V}(\epsilon) = \sigma^2$.

the index data from the Breeding Bird Survey over the years of study. Instead, Flather wanted an approximating model that fit the data well and could be used in making inferences about bird communities on the scale of large landscapes. In this first example, the number of parameters in the candidate models ranges only from 3 to 5. Which approximating model is "best" for making inferences from these data is answered philosophically by the principle of parsimony (Section 1.4) and operationally by several information-theoretic criteria in Chapter 2. Methods for estimating model selection uncertainty and incorporating this into inferences are given in Chapter 2 and illustrated in Chapters 4 and 5.

Note, in each case, that the response variable y is being modeled, rather than mixing models of y with $\log(y)$, or other transformations of the response variable (Table 1.1). These models are in the sense of 1 above, as the structure is given but the parameter values are unspecified. Given appropriate data, ML can be used to obtain $\hat{\theta}$ in the sense of 3 above. In some of the physical sciences the model parameters are derived from theory, without the need for problem-specific empirical data. Such cases seem to be the exception in the biological sciences, where model parameters must usually be estimated from the data using least squares or likelihood theory.

1.2.4 Science Inputs: Formulation of the Set of Candidate Models

Model specification or formulation, in its widest sense, is conceptually more difficult than estimating the model parameters and their precision. Model for-

mulation is the point where the scientific and biological information formally enter the investigation. Building the set of candidate models is partially a subjective art; that is why scientists must be trained, educated, and experienced in their discipline. The published literature and experience in the biological sciences can be used to help formulate a set of a priori candidate models. The most original, innovative part of scientific work is the phase leading to the proper question. Good approximating models, each representing a scientific hypothesis, in conjunction with a good set of relevant data can provide insight into the underlying biological process and structure.

Lehmann (1990) asks, "where do models come from," and cites some biological examples (also see Ludwig 1989, Walters 1996, Lindsey 1995). Models arise from questions about biology and the manner in which biological systems function. Relevant theoretical and practical questions arise from a wide variety of sources (see Box et al. 1978, O'Connor and Spotila 1992). Traditionally, these questions come from the scientific literature, results of manipulative experiments, personal experience, or contemporary debate within the scientific community. More practical questions stem from resource management controversies, biomonitoring programs, quasi-experiments, and even judicial hearings.

Chatfield (1995b) suggests that there is a need for more *careful thinking* (than is usually evident) and a *better balance* between the problem (biological question), analysis theory, and data. This suggestion has been made in the literature for decades. One must conclude that it has not been taught sufficiently in applied science or statistics courses. **Our science culture does not regularly do enough to expect and enforce critical thinking.** Too often, the emphasis is focused on the analysis theory and data analysis, with too little thought about the reason for the study in the first place (see Hayne 1978 for convincing examples).

Tukey (1980) argues for the need for deep thinking and early exploratory data analysis, and that the results of these activities lead to good scientific questions and confirmatory data analysis. In the exploratory phases, he suggests the importance of a flexible attitude and plotting of the data. He does not advocate the computation of test statistics, P-values, and so forth during exploratory data analysis. Tukey concludes that to implement the confirmatory paradigm properly we need to do a lot of exploratory work.

The philosophy and theory presented here must rest on well-designed studies and careful planning and execution of field or laboratory protocol. Many good books exist giving information on these important issues (Burnham et al. 1987, Cook and Campbell 1979, Mead 1988, Hairston 1989, Desu and Roghavarao 1991, Eberhardt and Thomas 1991, Manly 1992, Skalski and Robson 1992, Thompson 1992, Scheiner and Gurevitch 1993, Cox and Reid 2000, and Guisan and Zimmermann 2000). Chatfield (1991) reviews statistical pitfalls and ways that these might be avoided. Research workers are urged to pay close attention to these critical issues. Methods given here should not be thought to salvage poorly designed work. In the following material we will assume that the data

are "sound" and that inference to some larger population is reasonably justified by the manner in which the data were collected.

Development of the a priori set of candidate models often should include a global model: a model that has many parameters, includes all potentially relevant effects, and reflects causal mechanisms thought likely, based on *the science of the situation*. The global model should also reflect the study design and attributes of the system studied. Specification of the global model should not be based on a probing examination of the data to be analyzed. At some early point, one should investigate the fit of the global model to the data (e.g., examine residuals and measures of fit such as R^2, deviance, or formal χ^2 goodness-of-fit tests) and proceed with analysis only if it is judged that the global model provides an acceptable fit to the data. Models with fewer parameters can then be derived as special cases of the global model. This set of reduced models represents plausible alternatives based on what is known or hypothesized about the process under study. Generally, alternative models will involve differing numbers of parameters; the number of parameters will often differ by at least an order of magnitude across the set of candidate models. Chatfield (1995b) writes concerning the importance of subject-matter considerations such as accepted theory, expert background knowledge, and prior information in addition to known constraints on both the model parameters and the variables in the models. All these factors should be brought to bear on the makeup of the set of candidate models, prior to actual data analysis.

The more parameters used, the better the fit of the model to the data that is achieved. Large and extensive data sets are likely to support more complexity, and this should be considered in the development of the set of candidate models. **If a particular model (parametrization) does not make biological sense, this is reason to exclude it from the set of candidate models, particularly in the case where causation is of interest**. In developing the set of candidate models, one must recognize a certain balance between keeping the set small and focused on plausible hypotheses, while making it big enough to guard against omitting a very good a priori model. While this balance should be considered, we advise the inclusion of all models that seem to have a reasonable justification, prior to data analysis. While one must worry about errors due to both underfitting and overfitting, it seems that modest overfitting is less damaging than underfitting (Shibata 1989). We recommend and encourage a considerable amount of careful, a priori thinking in arriving at a set of candidate models (see Peirce 1955, Burnham and Anderson 1992, Chatfield 1995b).

Freedman (1983) noted that when there are many, say 50, explanatory variables ($x_1, x_2, \ldots, x_{50}$) used to predict a response variable (y), variable-selection methods will provide regression equations with high R^2 values, "significant" F values, and many "significant" regression coefficients, as shown by large t values, *even if the explanatory variables are independent of* y. This undesirable situation occurs most frequently when the number of variables is of the same order as the number of observations. This finding, known as Freedman's paradox, was illustrated by Freedman using hypothe-

sis testing as a means to select a model of y as a function of the x's, but the same type of problematic result can be found in using other model selection methods. Miller (1990) notes that estimated regression coefficients are biased away from zero in such cases; this is a type of model selection bias. The partial resolution of this paradox is in the a priori modeling considerations, keeping the number of candidate models small, achieving a large sample size relative to the number of parameters to be estimated, and basing inference on more than one model.

It is not uncommon to see biologists collect data on 50–130 "ecological" variables in the blind hope that some analysis method and computer system will "find the variables that are significant" and sort out the "interesting" results (Olden and Jackson 2000). This shotgun strategy will likely uncover mainly spurious correlations (Anderson et al. 2001b), and it is prevalent in the naive use of many of the traditional multivariate analysis methods (e.g., principal components, stepwise discriminant function analysis, canonical correlation methods, and factor analysis) found in the biological literature. We believe that mostly spurious results will be found using this unthinking approach (also see Flack and Chang 1987 and Miller 1990), and we encourage investigators to give very serious consideration to a well-founded set of candidate models and predictor variables (as a reduced set of possible prediction) as a means of minimizing the inclusion of spurious variables and relationships. Ecologists are not alone in collecting a small amount of data on a very large number of variables. A. J. Miller (personal communication) indicates that he has seen data sets in other fields with as many as 1,500 variables where the number of cases is less than 40 (a purely statistical search for meaningful relationships in such data is doomed to failure).

After a carefully defined set of candidate models has been developed, one is left with the evidence contained in the data; the task of the analyst is to interpret this evidence from analyzing the data. Questions such as, "What effects are supported by the data?" can be answered objectively. This modeling approach allows a clear place for experience (i.e., prior knowledge and beliefs), the results of past studies, the biological literature, and current hypotheses to enter the modeling process formally. Then, one turns to the data to see "what is important" within a sense of parsimony. In some cases, careful consideration of the number and nature of the predictor variables to be used in the analysis will suffice in defining the candidate models. This process may result in an initial set of, say, 15–40 predictor variables and a consolidation to a much smaller set to use in the set of candidate models. Using AIC and other similar methods one can only hope to select the best model from this set; if good models are not in the set of candidates, they cannot be discovered by model selection (i.e., data analysis) algorithms.

We lament the practice of generating models (i.e., "modeling") that is done in the total absence of real data, and yet "inferences" are made about the status, structure, and functioning of the real world based on studying these models. We do not object to the often challenging and stimulating intellectual exercise

of model construction as a means to integrate and explore our myriad ideas about various subjects. For example, Berryman et al. (1995) provide a nice list of 26 candidate models for predator–prey relationships and are interested in their "credibility" and "parsimony." However, as is often the case, there are no empirical data available on a variety of taxa to pursue these issues in a rigorous manner (also see Turchin and Batzli (2001), who suggest 8 models, each a system of 2–3 differential equations, for vegetation–herbivore population interactions). Such exercises help us sort out ideas that in fact conflict when their logical consequences are explored. Modeling exercises can strengthen our logical and quantitative abilities. Modeling exercises can give us insights into how the world *might* function, and hence modeling efforts can lead to alternative hypotheses to be explored with real data. Our objection is only to the confusing of presumed insights from such models with inferences about the real world (see Peters 1991, Weiner 1995). An inference from a model to some aspect of the real world is justified only after the model has been shown to adequately fit relevant empirical data (this will certainly be the case when the model in its totality has been fit to and tested against reliable data). Gause (1934) had similar beliefs when he stated, "Mathematical investigations independent of experiments are of but small importance . . . ."

The underlying philosophy of analysis is important here. We advocate a conservative approach to the overall issue of *strategy* in the analysis of data in the biological sciences with an emphasis on a priori considerations and models to be considered. *Careful, a priori consideration of alternative models will often require a major change in emphasis among many people*. This is often an unfamiliar concept to both biologists and statisticians, where there has been a tendency to use either a traditional model or a model with associated computer software, making its use easy (Lunneborg 1994). This a priori strategy is in contrast to strategies advocated by others who view modeling and data analysis as a highly iterative and interactive exercise. Such a strategy, to us, represents deliberate data dredging and should be reserved for early exploratory phases of initial investigation. Such an exploratory avenue is not the subject of this book.

Here, we advocate the deliberate exercise of carefully developing a set of, say, 4–20 alternative models as potential approximations to the population-level information in the data available and the scientific question being addressed (Lytle 2002 provides an advanced example). Some practical problems might have as many as 70–100 or more models that one might want to consider. The number of candidate models is often larger with large data sets. We find that people tend to include many models that are far more general than the data could reasonably support (e.g., models with several interaction parameters). There need to be some well-supported guidelines on this issue to help analysts better define the models to be considered. This set of models, developed without first deeply examining the data, constitutes the "*set of candidate models*." The science of the issue enters the analysis through the a priori set of candidate models.

1.2.5 Models Versus Full Reality

Fundamental to our paradigm is that none of the models considered as the basis for data analysis are the "true model" that generates the biological data we observe (see, for example, Bancroft and Han 1977). We believe that "truth" (full reality) in the biological sciences has essentially infinite dimension, and hence full reality cannot be revealed with only finite samples of data and a "model" of those data. It is generally a mistake to believe that there is a simple "true model" in the biological sciences and that during data analysis this model can be uncovered and its parameters estimated. Instead, biological systems are complex, with many small effects, interactions, individual heterogeneity, and individual and environmental covariates (most being unknown to us); we can only hope to identify a model that provides a good *approximation* to the data available. The words "true model" represent an oxymoron, except in the case of Monte Carlo studies, whereby a model is used to generate "data" using pseudorandom numbers (we will use the term "generating model" for such computer-based studies). The concept of a "true model" in biology seems of little utility and may even be a source of confusion about the nature of approximating models (e.g., see material on BIC and related criteria in Chapter 6).

A model is a simplification or approximation of reality and hence will not reflect all of reality. Taub (1993) suggests that unproductive debate concerning true models can be avoided by simply recognizing that a model is not truth by definition. Box (1976) noted that "all models are wrong, but some are useful." While a model can never be "truth," a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless. Model selection methods try to rank models in the candidate set relative to each other; whether any of the models is actually "good" depends primarily on the quality of the data and the science and a priori thinking that went into the modeling. Full truth (reality) is elusive (see deLeeuw 1988). Proper modeling and data analysis tell what inferences the data support, not what full reality might be (White et al. 1982:14–15, Lindley 1986). Models, used cautiously, tell us "what effects are supported by the (finite) data available." Increased sample size (information) allows us to chase full reality, but never quite catch it.

The concept of truth and the false concept of a true model are deep and surprisingly important. Often, in the literature, one sees the words *correct* model or simply *the* model as if to be vague as to the exact meaning intended. Bayesians seem to say little about the subject, even as to the exact meaning of the prior probabilities on models. Consider the simple model of population size (n) at time t,

$$n_{t+1} = n_t \cdot s_t,$$

where s is the survival probability during the interval from t to $t + 1$. This is a correct model in the sense that it is algebraically and deterministically correct; however, it is not an exact representation or model of truth. This model is not explanatory; it is definitional (it is a tautology, because it implies that

$s_t = n_{t+1}/n_t$). For example, from the theory of natural selection, the survival probability differs among the n animals. Perhaps the model above could be improved if average population survival probability was a random variable from a beta distribution; still, this is far from a model of full reality or truth, even in this very simple setting. Individual variation in survival could be caused by biotic and abiotic variables in the environment. Thus, a more exact model of full reality would have, at the very least, the survival of each individual as a nonlinear function of a large number of environmental variables and their interaction terms. Even in this simple case, it is surely clear that one cannot expect any mathematical model to represent full reality; there are no true models in the biological sciences. We will take a set of approximating models g_i, without pretending that one represents full reality and is therefore "true."

In using some model selection methods it is assumed that the set of candidate models contains the "true model" that generated the data. We will not make this assumption, unless we use a data set generated by Monte Carlo methods as a tutorial example (e.g., Section 3.4), and then we will make this artificial condition clear. In the analysis of real data, it seems unwarranted to pretend that the "true model" is included in the set of candidate models, or even that the true model exists at all. Even if a "true model" did exist and if it could be found using some method, it would not be good as a fitted model for general inference (i.e., understanding or prediction) about some biological system, because its numerous parameters would have to be estimated from the finite data, and the precision of these estimated parameters would be quite low.

Often the investigator wants to simplify some representation of reality in order to achieve an understanding of the dominant aspects of the system under study. If we were given a nonlinear formula with 200 parameter values, we could make correct predictions, but it would be difficult to *understand* the main dynamics of the system without some further simplification or analysis. Thus, one should tolerate some inexactness (an inflated error term) to facilitate a simpler and more useful understanding of the phenomenon.

In particular, we believe that there are tapering effect sizes in many biological systems; that is, there are often several large, important effects, followed by many smaller effects, and, finally, followed by a myriad of yet smaller effects. These effects may be sequentially unveiled as sample size increases. The main, dominant, effects might be relatively easy to identify and support, even using fairly poor analysis methods, while the second-order effects (e.g., a chronic treatment effect or an interaction term) might be more difficult to detect. The still smaller effects can be detected only with very large sample sizes (cf. Kareiva 1994 and related papers), while the smallest effects have little chance of being detected, even with very large samples. Rare events that have large effects may be very important but quite difficult to study. Approximating models must be related to the amount of data and information available; small data sets will appropriately support only simple models with few parameters, while more comprehensive data sets will support, if necessary, more complex models.

This tapering in "effect size" and high dimensionality in biological systems might be quite different from some physical systems where a small-dimensioned model with relatively few parameters might accurately represent full truth or reality. Biologists should not believe that a simple "true model" exists that generates the data observed, although some biological questions might be of relatively low dimension and could be well approximated using a fairly simple model. The issue of a range of tapering effects has been realized in epidemiology, where Michael Thun notes, "... you can tell a little thing from a big thing. What's very hard to do is to tell a little thing from nothing at all" (Taubes 1995). *Full reality will always remain elusive in the biological sciences*.

At a more advanced conceptual level, these is a concept that "information" about the population (or process or system) under study exists in the data and the goal is to express this information in a more compact, understandable form using a "model." Conceptually, this is a change in coding system, similar to using a different "alphabet." The data have only a finite, fixed amount of information. The *goal* of model selection is to achieve a perfect one-to-one translation so that no information is lost; in fact, we cannot achieve this ideal. The data can be ideally partitioned into *information* and *noise*. The noise part of the data is not information. However, noise could contain information that we cannot decode. Conceptually, the role of a good model is to filter the data so as to separate information from noise.

Our main emphasis in modeling empirical data is to understand the biological structure, process, or system. Sometimes prediction will be of interest; here, however, one would hopefully have an understanding of the structure of the system as a basis for making trustworthy predictions. We recommend developing a set of candidate models prior to intensive data analysis, selecting one that is "best," and estimating the parameters of that model and their precision (using maximum likelihood or least squares methods). This unified strategy is a basis for valid inferences, and there are several more advanced methods to allow additional inferences and insights. In particular, models exist to allow formal inference from more than one model, and this has a number of advantages (Hoeting et al. 1999). Statistical science is not so much a branch of mathematics, but rather it is concerned with the development of a practical theory of information using what is known or postulated about the science of the matter. In our investigations into these issues we were often surprised by how much uncertainty there is in selecting a good approximating model; the variability in terms of what model is selected or considered best from independent data sets, for example, is often large.

1.2.6 An Ideal Approximating Model

We consider some properties of an ideal model for valid inference in the analysis of data. It is important that the best model is selected from a set of models that were defined prior to data analysis and based on the science of the issue

at hand. Ideally, the process by which a "best" model is selected would be objective and repeatable; these are fundamental tenets of science. The ideal model would be appropriately simple, based on concepts of parsimony. Furthermore, precise, unbiased estimators of parameters would be ideal, as would accurate estimators of precision. The best model would ideally yield achieved confidence interval coverage close to the nominal level (often 0.95) and have confidence intervals of minimum width. Achieved confidence interval coverage is a convenient index to whether parameter estimators and measures of precision are adequate. Finally, one would like as good an approximation of the structure of the system as the information permits. Thus, in many cases adjusted R^2 can be computed and σ^2 estimated as a measure of variation explained or residual variation, respectively. Ideally, the parameters in the best model would have biological interpretations. If prediction was the goal, then having the above issues in place might warrant some tentative trust in model predictions. There are many cases where two or more models are essentially tied for "best," and this should be fully recognized in further analysis and inference, especially when they produce different predictions. In other cases there might be 4–10 models that have at least some support, and these, too, deserve scrutiny in reaching conclusions from the data, based on inferences from more than a single model.

1.3 Model Fundamentals and Notation

This section provides a conceptualization of some important classes of models as they are used in this book. Some of these classes are particularly important in model selection. A general notation is introduced that is intended to be helpful to readers.

1.3.1 Truth or Full Reality f

While there are no models that exactly represent full reality (cf. Section 1.2.5), full truth can be denoted as f. The concept of f is abstract. It is this truth to which we want to make inferences, based on data and approximating models. We use the notation $f(x)$ to denote that integration is over the variable x, but we do not want to convey the notion that f is a function of the data x. Data arise from full reality and can be used to make formal inferences back to this truth, if data collection has been carefully planned and proper sampling or experimental design has been achieved.

1.3.2 Approximating Models $g_i(x|\theta)$

We use the notation $g_i(x|\theta)$ or often, if the context is clear, g_i to denote the ith approximating model. We use θ to represent generally a parameter or

vector of parameters. Thus, θ is generic and might represent parameters in a regression model ($\beta_0, \beta_1, \beta_2$) or the probability of a head in penny flipping trials (p). The models g_i are discrete or continuous probability distributions, and our focus will be on their associated likelihoods, $\mathcal{L}(\theta|data,\ model)$ or log-likelihoods $\log(\mathcal{L}(\theta|data,\ model))$. Notation for the log-likelihood will sometimes be shortened to $\log(\mathcal{L}(\theta|x, g))$ or even $\log(\mathcal{L})$. Ideally, the set of R models will have been defined prior to data analysis. These models specify only the form of the model, leaving the unknown parameters (θ) unspecified.

A simple example will aid in the understanding of this section. Consider a study of mortality (μ_c) as a function of concentration (c) of some chemical compound. The size (s) of the animal (binary as small or large) and a group covariate (z, such as gender) are also recorded, because they are hypothesized to be important in better understanding the concentration–mortality function. Investigators might consider mortality probability during some fixed time interval to be a logistic function of concentration, where, for example, $c = 0, 1, 2, 4, 8,$ and 16. The full structure of the logistic model when all 3 variables are included in the model can be written as,

$$\mu_c = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 c + \beta_2 s + \beta_3 z)\}}.$$

Use of the logistic link function allows the expression to be written as a linear model structure,

$$\text{logit}(\mu_c) = \log_e\left(\frac{\mu_c}{1-\mu_c}\right) = \beta_0 + \beta_1 c + \beta_2 s + \beta_3 z.$$

Here the data (y) are binary for mortality (dead or alive), size (small or large), and gender (male and female), while concentration is recorded at 6 fixed levels. The response variable $y = 1$ if the animal died and 0 if it lived, given a particular concentration. Then,

$$\text{Prob}\{y = 1|c, s, z\} = \mu_c$$

for n individuals at concentration c, size s, and gender z. Then, the likelihood is proportional to

$$\mathcal{L}(\mu_c|data,\ model) = \prod_{i=1}^{n} (\mu_c(i))^{y_i}\,(1-\mu_c(i))^{1-y_i}\ .$$

Thus, a set of approximating structural models might be defined, based on the science of the issue. The stochastic part of the model is assumed to be Bernoulli. The models are alternatives, defined prior to data analysis, and the interest is in the strength of evidence for each of the alternative hypotheses, represented by models. Five ($R = 5$) structural models will be used for illustration:

$$\begin{aligned}
g_1(x):&\ \ \text{logit}(\mu_c) = \beta_0 + \beta_1 c + \beta_2 s + \beta_3 z,\\
g_2(x):&\ \ \text{logit}(\mu_c) = \beta_0 + \beta_1 c + \beta_2 s,\\
g_3(x):&\ \ \text{logit}(\mu_c) = \beta_0 + \beta_1 c \qquad\quad + \beta_3 z,
\end{aligned}$$

$$g_4(x): \quad \text{logit}(\mu_c) = \beta_0 + \beta_1 c,$$
$$g_5(x): \quad \text{logit}(\mu_c) = \beta_0.$$

These models specify the structural form (including how the parameters and covariates enter), but not the parameter values (the β_i); each assumes that the y are independent Bernoulli random variables. The first model serves as a global model. The second model represents the hypothesis that the group covariate (z) is unimportant, while the third model is like the first, except that the size is hypothesized to be unimportant. The fifth model implies that mortality is constant and not a function of concentration. Often, enough is known about the compound that model g_5 is not worth exploration. Of course, the log-log or complementary log-log, or probit function could have been used to model the hypothesized relationships in this example, rather than the logistic.

1.3.3 The Kullback–Leibler Best Model $g_i(x|\theta_0)$

For given full reality (f), data (x), sample size (n), and model set (R) there is a best model in the sense of Kullback-Leibler information (introduced in Chapter 2). That is, given the possible data, the form of each model, and the possible parameter values, K-L information can be computed for each model in the set and the model best approximating full reality determined.

The parameters that produce this conceptually best single model, in the class $g(x|\theta)$, are denoted by θ_0, Of course, this model is generally unknown to us but can be estimated; such estimation involves computing the MLEs of the parameters in each model ($\hat{\theta}$) and then *estimating* K-L information as a basis for model selection and inference. The MLEs converge asymptotically to θ_0 and the concept of bias is with respect to θ_0, rather than our conceptual "true parameters" associated with full reality f.

1.3.4 Estimated Models $g_i(x|\hat{\theta})$

Estimated models have specific parameter values from ML or LS estimation, based on the given data and model. If another, replicate data set were available and based on the same sample size, the parameter estimates would differ somewhat; the amount of difference expected is related to measures of precision (e.g., standard errors and confidence intervals). It is important to keep separate the model form $g_i(x|\theta)$ from specific estimates of this model, based on data and the process of parameter estimation, $g_i(x|\hat{\theta})$.

In the models of mortality as a function of concentration and other variables (above), there are associated likelihoods and log-likelihoods. Likelihood theory can be used to obtain the MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$ for model g_4, for example. The likelihood function is

$$\mathcal{L}(\beta_0, \beta_1 | \textit{data}, \ \textit{model}) = \prod_{i=1}^{n} (\mu_c(i))^{y_i} \, (1 - \mu_c(i))^{1-y_i} \, ,$$

where

$$\mu_c = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 c)\}}.$$

Thus, the only parameters in the likelihood are β_0 and β_1 and given the data, one can obtain the MLEs. The value of the maximized log-likelihood and the estimated variance–covariance matrix can also be computed. In a sense, when we have only the model form $g(x|\theta)$ we have an infinite number of models, where all such models have the same form but different values of θ. Yet, in all of these models there is a unique K-L best model. Conceptually, we know how to find this model, given f.

1.3.5 Generating Models

Monte Carlo simulation is a very useful and general approach in theoretical and applied statistics (Manly 1991). These procedures require that a model be specified as the basis for generating Monte Carlo data. Such a model is not full reality, and thus we call it a *generating model*. It is "truth" only in the sense of computerized truth. One should not confuse a generating model or results based on Monte Carlo data with full reality f.

1.3.6 Global Model

Ideally, the global model has in it all the factors or variables thought to be important. Other models are often special cases of this global model. There is not always a global model. If sample size is small, it may be impossible to fit the global model. Goodness-of-fit tests and estimates of an overdispersion parameter for count data should be based (only) on the global model. The concept of overdispersion is relatively model-independent; however, some model must be used to compute or model any overdispersion thought to exist in count data. Thus, the most highly parametrized model will serve best as the basis for assessing overall fit and estimating a parameter associated with overdispersion. In the models of mortality (above), model g_1 would serve as the global model.

The advantage of this approach is that if the global model fits the data adequately, then a selected model that is more parsimonious will also fit the data (this is an empirical result, not a theorem). Parsimonious model selection should not lead to a model that does not fit the data (this property seems to hold for the selection methods we advocate here). Thus, goodness-of-fit assessment and the estimation of overdispersion parameters should be addressed using the global model (this could also be computed for the selected model).

In summary, we will use the word "model" to mean different things; hopefully, the context will be clear. Certainly it is important to distinguish clearly between f and g. The general structural form is denoted by $g(x|\theta)$, without specifying the numerical value of the parameter θ (e.g., models given in Table

1.1). If one considers estimation of θ, then there are an infinite number of possible values of θ. Therefore, there is an entire class of models $g(x|\theta)$, defined by the space over which θ varies. Frequently, we will refer to the model where MLEs (the most likely, given the data and the model) have been found. In other cases we will mean the best model, $g(x|\theta_0)$, which is one specific model (the K-L best relative to f).

1.3.7 Overview of Stochastic Models in the Biological Sciences

Models are useful in the biological sciences for understanding the structure of systems, estimating parameters of interest and their associated variance–covariance matrix, predicting outcomes and responses, and testing scientific hypotheses. Such models might be used for "relational" or "explanatory" purposes or might be used for prediction. In the following material we will review the main types of models used in the biological sciences. Although the list is not meant to be exhaustive, it will allow the reader an impression of the wide class of models of empirical data that we will treat under an information-theoretic framework.

Simple linear and multiple linear regression models (Seber 1977, Draper and Smith 1981, Brown 1993) have seen heavy use in the biological sciences over the past four decades. These models commonly employ one to perhaps 8–12 parameters, and the statistical theory is fully developed (either based on least squares or likelihood theory). Similarly, analysis of variance and covariance models have been widely used, and the theory underlying these methods is closely related to regression models and is fully developed (both are examples of general linear models). Theory and software for this wide class of methods are readily available.

Nonlinear regression models (Gallant 1987, Seber and Wild 1989, Carroll et al. 1995) have also seen abundant use in the biological sciences (logistic regression is a common example). Here, the underlying theory is often likelihood based, and some classes of nonlinear models require very specialized software. In general, nonlinear estimation is a more advanced problem and is somewhat less well understood by many practicing researchers.

Other types of models used in the biological sciences include generalized linear (McCullagh and Nelder 1989, Morgan 1992, 2000) and generalized additive (Hastie and Tibshirani 1990) models (these can be types of nonlinear regression models). These modeling techniques have seen increasing use in the past decade. Multivariate modeling approaches such as multivariate ANOVA and regression, canonical correlation, factor analysis, principal components analysis, and discriminate function analysis have had a checkered history in the biological and social sciences, but still see substantial use (see review by James and McCulloch 1990). Log-linear and logistic models (Agresti 1990) have become widely used for count data. Time series models (Brockwell and Davis 1987, 1991) are used in many biological disciplines. Various models of an organism's growth (Brisbin et al. 1987, Gochfeld 1987) have been proposed and

used in biology. Caswell (2001) provides a large number of matrix population models that have seen wide use in the biological sciences.

Compartmental models are a type of state transition in continuous time and continuous response and are usually based on systems of differential or partial differential equations (Brown and Rothery 1993, Matis and Kiffe 2000). There are discrete state transition models using the theory of Markov chains (Howard 1971); these have found use in a wide variety of fields including epidemiological models of disease transmission. More advanced methods with potentially wide application include the class of models called "random effects" (Kreft and deLeeuw 1998).

Models to predict population viability (Boyce 1992), often based on some type of Leslie matrix, are much used in conservation biology, but rarely are alternative model forms given serious evaluation. A common problem here is that these models are rarely based on empirical data; the form of the model and its parameter values are often merely only "very rough guesses" necessitated by the lack of empirical data (White 2000).

Biologists in several disciplines employ differential equation models in their research (see Pascual and Kareiva 1996 for a reanalysis of Gause's competition data and Roughgarden 1979 for examples in population genetics and evolutionary ecology). Many important applications involve exploited fish populations (Myers et al. 1995). Computer software exists to allow model parameters to be estimated using least squares or maximum likelihood methods (e.g., SAS and Splus). These are powerful tools in the analysis of empirical data, but also beg the issue of "what model to use."

Open and closed capture–recapture (Lebreton et al. 1992) and band recovery (Brownie et al. 1985) models represent a class of models based on product multinomial distributions (see issues 5 and 6 of volume 22 of the *Journal of Applied Statistics*, 1995). Distance sampling theory (Buckland et al. 1993, 2001) relies on models of the detection function and often employs semiparametric models. Parameters in these models are nearly always estimated using maximum likelihood.

Spatial models (Cressie 1991 and Renshaw 1991) are now widely used in the biological sciences, allowing the biologist to take advantage of spatial data sets (e.g., geographic information systems). Stein and Corsten (1991) have shown how Kriging (perhaps the most widely used spatial technique) can be expressed as a least squares problem, and the development of Markov chain Monte Carlo methods such as the Gibbs sampler (Robert and Casella 1999, Chen et al. 2000) allow other forms of spatial models to be fitted by least squares or maximum likelihood (Augustin et al. 1996). Further unifying work for methods widely used on biological data has been carried out by Stone and Brooks (1990). Geographic information systems potentially provide large numbers of covariates for biological models, so that model selection issues are particularly important.

Spatiotemporal models are potentially invaluable to the biologist, though most researchers model changes over space or time, and not both simultane-

ously. The advent of Markov chain Monte Carlo methods (Gilks et al. 1996, Gamerman 1997) may soon give rise to a general but practical framework for spatiotemporal modeling; model selection will be an important component of such a framework. A step towards this general framework was made by Buckland and Elston (1993), who modeled changes in the spatial distribution of wildlife.

There are many other examples where modeling of data plays a fundamental role in the biological sciences. Henceforth, we will exclude only modeling that cannot be put into a likelihood or quasi-likelihood (Wedderburn 1974) framework and models that do not explicitly relate to empirical data. All least squares formulations are merely special cases that have an equivalent likelihood formulation in usual practice. There are general information-theoretic approaches for models well outside the likelihood framework (Qin and Lawless 1994, Ishiguo et al. 1997, Hurvich and Simonoff 1998, and Pan 2001a and b). There are now model selection methods for nonparametric regression, splines, kernel methods, martingales, and generalized estimation equations. Thus, methods exist for nearly all classes of models we might expect to see in the theoretical or applied biological sciences.

1.4 Inference and the Principle of Parsimony

1.4.1 Avoid Overfitting to Achieve a Good Model Fit

Consider two analysts studying a small set of biological data using a multiple linear regression model. The first exclaims that a particular model provides an excellent fit to the data. The second notices that 22 parameters were used in the regression and states, "Yes, but you have used enough parameters to fit an elephant!" This seeming conflict between increasing model fit and increasing numbers of parameters to be estimated from the data led Wel (1975) to answer the question, "How many parameters *does* it take to fit an elephant?" Wel finds that about 30 parameters would do reasonably well (Figure 1.2); of course, had he fit 36 parameters to his data, he could have achieved a perfect fit.

Wel's finding is both insightful and humorous, but it deserves further interpretation for our purposes here. His "standard" is itself only a crude drawing—it even lacks ears, a prominent elephantine feature; hardly truth. A better target would have been a large, digitized, high-resolution photograph; however, this, too, would have been only a model (and not truth). Perhaps a real elephant should have been used as truth, but this begs the question, "Which elephant should we use?" This simple example will encourage thinking about full reality, "true models," and approximating models and motivate the *principle of parsimony* in the following section. **William of Occam suggested in the fourteenth century that one "shave away all that is unnecessary"—a dictum often referred to as *Occam's razor*. Occam's razor has had a long history**

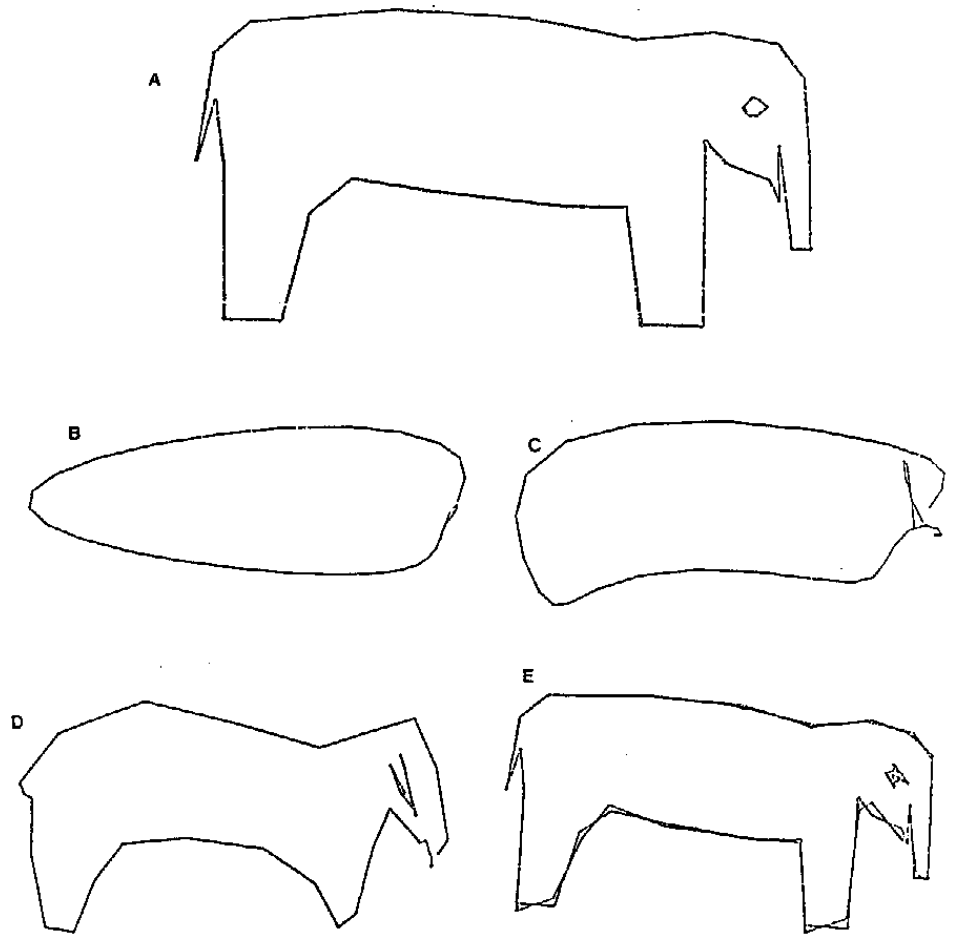

FIGURE 1.2. "How many parameters does does it take to fit an elephant?" was answered by Wel (1975). He started with an idealized drawing (A) defined by 36 points and used least squares Fourier sine series fits of the form $x(t) = \alpha_0 + \sum \alpha_i \sin(it\pi/36)$ and $y(t) = \beta_0 + \sum \beta_i \sin(it\pi/36)$ for $i = 1, \ldots, N$. He examined fits for $K = 5, 10, 20$, and 30 (shown in B–E) and stopped with the fit of a 30 term model. He concluded that the 30-term model "may not satisfy the third-grade art teacher, but would carry most chemical engineers into preliminary design."

in both science and technology, and it is embodied in the principle of parsimony. Albert Einstein is supposed to have said, "Everything should be made as simple as possible, but no simpler."

Success in the analysis of real data and the resulting inference often depends importantly on the choice of a best approximating model. Data analysis in the biological sciences should be based on a parsimonious model that provides an accurate approximation to the structural information in the data at hand; this should not be viewed as searching for the "true model." Modeling and model selection are essentially concerned with the "art of approximation" (Akaike 1974).

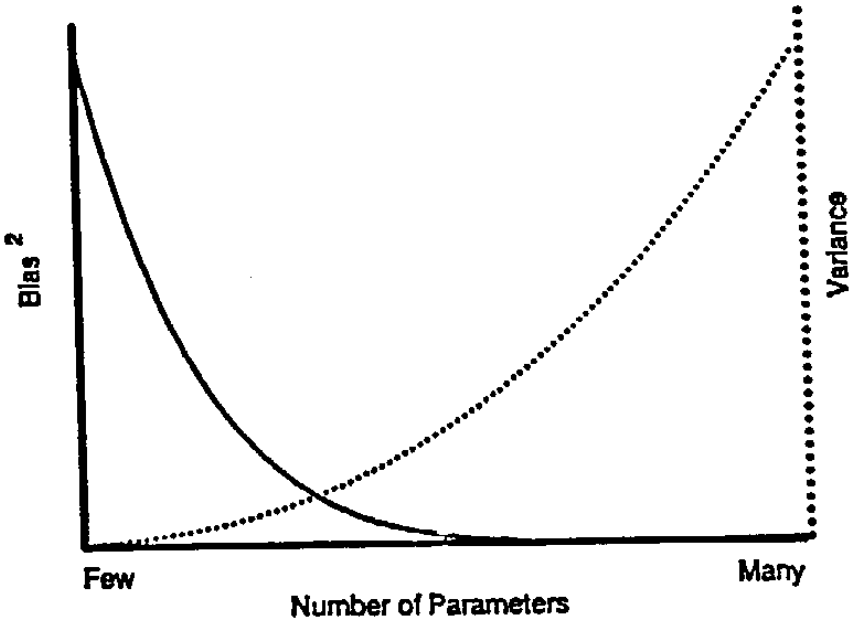

FIGURE 1.3. The *principle of parsimony*: the conceptual tradeoff between squared bias (solid line) and variance vs. the number of estimable parameters in the model (K). All model selection methods implicitly employ some notion of this tradeoff. The best approximating model need not occur exactly where the two curves intersect. Full truth or reality is not attainable with finite samples and usually lies well to the right of the region in which the best approximating model lies (the tradeoff region). Bias decreases and variance (uncertainty) increases as the number of parameters in a model increases.

1.4.2 The Principle of Parsimony

If the fit is improved by a model with more parameters, then where should one stop? Box and Jenkins (1970:17) suggested that the *principle of parsimony* should lead to a model with "... the smallest possible number of parameters for adequate representation of the data." Statisticians view the principle of parsimony as a bias versus variance tradeoff. In general, bias decreases and variance increases as the dimension of the model (K) increases (Figure 1.3). Often, we may use the number of parameters in a model as a measure of the degree of structure inferred from the data. The fit of any model can be improved by increasing the number of parameters (e.g., the elephant-fitting problem); however, a tradeoff with the increasing variance must be considered in selecting a model for inference. Parsimonious models achieve a proper tradeoff between bias and variance. All model selection methods are based to some extent on the principle of parsimony (Breiman 1992, Zhang 1994).

In understanding the utility of an approximate model for a given data set, it is convenient to consider two undesirable possibilities: underfitted and overfitted models. Here, we must avoid judging a selected model in terms of some supposed "true model," as occurs when data are simulated from a known, often very simple, model using Monte Carlo methods. In this case, if the generating

model had 10 parameters, it is often said that an approximating model with only 7 parameters is underfitted (compared with the generating model with 10 parameters). This interpretation is often of little value, because it largely ignores the principle of parsimony and its implications and hinges on the misconception that such a simple true model exists in biological problems. If we believe that truth is essentially infinite-dimensional, then overfitting is not even defined in terms of the number of parameters in the fitted model. We will avoid this use of the terms "underfitted" and "overfitted" that suppose the existence of a low-dimensional "true model" as a "standard."

Instead, we reserve the terms underfitted and overfitted for use in relation to a "best approximating model" (Section 1.2.6). Here, an underfitted model would ignore some important replicable (i.e., conceptually replicable in most other samples) structure in the data and thus fail to identify effects that were actually supported by the data. In this case, bias in the parameter estimators is often substantial, and the sampling variance is underestimated, both factors resulting in poor confidence interval coverage. Underfitted models tend to miss important treatment effects in experimental settings. Overfitted models, as judged against a best approximating model, are often free of bias in the parameter estimators, but have estimated (and actual) sampling variances that are needlessly large (the precision of the estimators is poor, relative to what could have been accomplished with a more parsimonious model). Spurious treatment effects tend to be identified, and spurious variables are included with overfitted models. Shibata (1989) argues that underfitted models are a more serious issue in data analysis and inference than overfitted models. This assessment breaks down in many exploratory studies where sample size might be only 35–80 and there are 20–80 explanatory variables. In these cases, one may expect substantial overfitting and many effects that are actually spurious (Freedman 1983, Anderson et al. 2001b).

The concept of parsimony and a bias versus variance tradeoff is very important. Thus we will provide some additional insights (also see Forster 1995, Forster and Sober 1994, and Jaffe and Spirer 1987). The goal of data collection and analysis is to make inferences from the sample that properly apply to the population. The inferences relate to the *information* about structure of the system under study as inferred from the models considered and the parameters estimated in each model. A paramount consideration is the repeatability, with good precision, of any inference reached. When we imagine many replicate samples, there will be some recognizable features common to almost all of the samples. Such features are the sort of inference about which we seek to make strong inferences (from our single sample). Other features might appear in, say, 60% of the samples yet still reflect something real about the population or process under study, and we would hope to make weaker inferences concerning these. Yet additional features appear in only a few samples, and these might be best included in the error term (σ^2) in modeling. If one were to make an inference about these features quite unique to just the single data set at hand, as if they applied to all (or most all) samples (hence to the population), then

we would say that the sample is overfitted by the model (we have overfitted the *data*). Conversely, failure to identify the features present that are strongly replicable over samples is underfitting. The data are not being approximated; rather we approximate the structural information in the data that is replicable over such samples (see Chatfield 1996, Collopy et al. 1994). Quantifying that structure with a model form and parameter estimates is subject to some "sampling variation" that must also be estimated (inferred) from the data.

True replication is very advantageous, but this tends to be possible only in the case of strict experiments where replication and randomization are a foundation. Such experimental replication allows a valid estimate of residual variation (σ^2). An understanding of these issues makes one realize what is lost when observational studies seem possible and practical, and strict experiments seem less feasible.

A best approximating model is achieved by properly balancing the errors of underfitting and overfitting. Stone and Brooks (1990) comment on the "... straddling pitfalls of underfitting and overfitting." The proper balance is achieved when bias and variance are controlled to achieve confidence interval coverage at approximately the nominal level and where interval width is at a minimum. Proper model selection rejects a model that is far from reality and attempts to identify a model in which the error of approximation and the error due to random fluctuations are well balanced (Shibata 1983, 1989). Some model selection methods are "parsimonious" (e.g., BIC, Schwarz 1978) but tend, in realistic situations, to select models that are too simple (i.e., underfitted); thus, bias is large, precision is overestimated, and achieved confidence interval coverage is well below the nominal level. Such instances are not satisfactory for inference. One has only a highly precise, quite biased result.

Sakamoto et al. (1986) simulated data to illustrate the concept of parsimony and the errors of underfitting and overfitting models (Figure 1.4). Ten data sets (each with $n = 21$) were generated from the simple model

$$y = e^{(x-0.3)^2} - 1 + \epsilon,$$

where x varied from 0 to 1 in equally spaced steps of 0.05, and $\epsilon \sim N(0, 0.01)$. Thus, in this case, they considered the generating model to have $K = 3$ parameters: 0.3, -1, and 0.01. They considered the set of candidate models (i.e., the approximating models) to be simple polynomials of order 0 to 5, as in the table below.

Order	K	Approximating Model
0	2	$\mathrm{E}(y) = \beta_0$
1	3	$\mathrm{E}(y) = \beta_0 + \beta_1(x)$
2	4	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2)$
3	5	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2) + \beta_3(x^3)$
4	6	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2) + \beta_3(x^3) + \beta_4(x^4)$
5	7	$\mathrm{E}(y) = \beta_0 + \beta_1(x) + \beta_2(x^2) + \beta_3(x^3) + \beta_4(x^4) + \beta_5(x^5)$.

Thus, each of these 6 models was fit to each of the 10 simulated data sets.

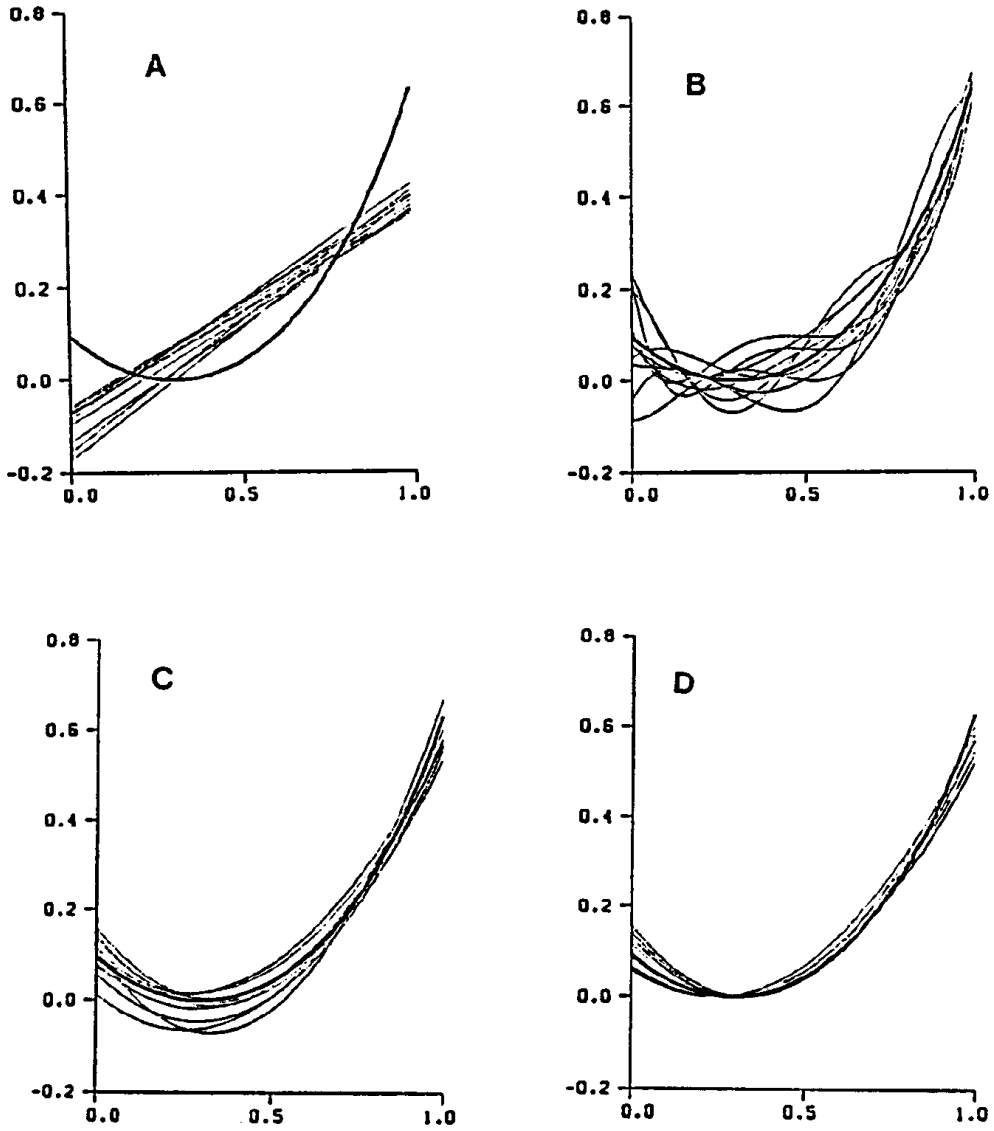

FIGURE 1.4. Ten Monte Carlo repetitions of data sets ($n = 21$) generated from the model $y = e^{(x-0.3)^2} - 1 + \epsilon; 0 \leq x \leq 1, \epsilon \sim N(0, .01)$ (from Sakamoto et al. 1986:164–179). A 1st-order polynomial (A) clearly misidentifies the basic nonlinear structure, and is underfitted and unsatisfactory. A 5th-order polynomial (B) has too many parameters, an unnecessarily large variance, and will have poor predictive qualities because it is unstable (overfitted). Neither A nor B is properly parsimonious, nor do they represent a best approximating model. A 2nd-order polynomial seems quite good as an approximating model (C). If it is known that the function is nonnegative and has its minimum at $x = 0.3$, then the approximating model that enforces these conditions is improved further (D). In more realistic situations, one lacks the benefit of simple plots and 10 independent data sets, such as those shown in A–D. See Section 3.7 for a full analysis of these data.

Strong model bias occurs when an underfitting (e.g., the mean-only model with $K = 2$ or the linear, 1st order, $K = 3$) model is employed (Figure 1.4A). Here bias is obvious, the nonlinear structure of the generating model is poorly approximated, and confidence interval coverage and predictions from the model will be quite poor. Of course, there is *some* model bias for each of the 5 models because they are only simple polynomial approximations. Overfitting is illustrated in Figure 1.4B, where a 5th-order polynomial ($K = 7$) is used as an approximating model. Here, there is little evidence of bias (an average quantity), precision is obviously poor, and it is difficult to identify the simple structure of the model. Prediction will be quite imprecise from this model, and it has features that do not occur in the generating model, particularly if one extrapolates beyond the range of the data (always a risky practice). Both underfitting and overfitting are undesirable in judging approximating models for data analysis.

If a second-order polynomial ($K = 4$) is used as the approximating model, the fits seem quite reasonable (Figure 1.4C), and one might expect valid inference from this model. Finally, *if* it were known a priori from the science of the situation that the function was nonnegative and had a minimum of zero at $x = 0.3$, then an improved quadratic approximating model could use this information very effectively (Figure 1.4D). The form of this model is

$$\mathrm{E}(y) = \beta_0(x + \beta_1)^2$$

with $K = 3$ (i.e., β_0, β_1, and σ^2), whereas the second-order polynomial has 4 parameters. This example illustrates that valid statistical inference is only partially dependent on the analysis process; the science of the situation must play an important role through modeling. This particular example provides a visual image of underfitting and overfitting in a simple case where the generating model and various approximating models can be easily graphed in two dimensions. Parsimony issues with real data in the biological sciences nearly always defy such a simple graphical approach because truth is not known; one rarely has 10 independent data sets on exactly the same process, and plots in high dimensions are problematic to produce and interpret. Note, also, that the generating model contained no tapering effects. However, the approximating models do have tapering effects. Therefore, objective and effective methods are needed that do not rely on simple graphics and can cope with the real-world complexities and high dimensionality.

1.4.3 Model Selection Methods

Model selection has most often been viewed, and hence taught, in a context of null hypothesis testing. Sequential testing has most often been employed, either stepup (forward) or stepdown (backward) methods. Stepwise procedures allow for variables to be added or deleted at each step. These testing-based methods remain popular in many computer software packages in spite of their poor operating characteristics. Testing schemes are based on subjective α levels;

commonly 0.05 or 0.01; however, Rawlings (1988) recommends 0.15 in the context of stepwise regression. The multiple testing problem is serious if many tests are to be made (see Westfall and Young 1993), and the tests are not independent. Tests between models that are not nested are problematic. A model is nested if it is a special case of another model; for example, a third-degree polynomial is nested within a fourth-degree polynomial. Generally, hypothesis testing is a very poor basis for model selection (Akaike 1974 and Sclove 1994b). McQuarrie and Tsai (1998) do not even treat this subject except for a short appendix on stepwise regression—the final three pages in their book.

Cross-validation has been suggested and well studied as a basis for model selection (Mosteller and Tukey 1968, Stone 1974, 1977; Geisser 1975). Here, the data are divided into two partitions. The first partition is used for model fitting; and the second is used for model validation (sometimes the second partition has only one observation). Then a new partition is selected, and this whole process is repeated hundreds or thousands of times. Some criterion is then chosen, such as minimum squared prediction error, as a basis for model selection. There are several variations on this theme, and it is a useful methodology (Craven and Wahba 1979, Burman 1989, Shao 1993, Zhang 1993a, and Hjorth 1994). These methods are quite computer intensive and tend to be impractical if more than about 15–20 models must be evaluated or if sample size is large. Still, cross-validation offers an interesting alternative for model selection.

Some analysts favor using a very general model in all cases (e.g., an overfitted model). We believe that this is generally poor practice (Figure 1.3B). Others have a "favorite" model that they believe is good, and they use it in nearly all situations. For example, some researchers always use the hazard rate model (Buckland et al. 1993) with 2 parameters ($K = 2$) as an approximating model to the detection function in line transect sampling. This might be somewhat reasonable for situations where a simple model suffices (e.g., $K = 2$ to 3), but will be poor practice in more challenging modeling contexts where $10 \leq K \leq 30$ or more is required. These *ad hoc* rules ignore the principle of parsimony and data-based model selection, in which the data help select the model to be used for inference.

If goodness-of-fit tests can be computed for all alternative models even if some are not nested within others, then one could use the model with the fewest parameters that "fits" (i.e., $P > 0.05$ or 0.10). However, increasingly better fits can often be achieved by using models with more and more parameters (e.g., the elephant-fitting problem), and this can make the arbitrary choice of α very critical. A large α-level leads to overfitted models and their resulting problems. In addition, other problems may be encountered such as over- or underdispersion and low power if one must pool small expectations to ensure that the test statistic is chi-square distributed. Perhaps, most importantly, there is no theory to suggest that this approach will lead to selected models with good inferential properties (i.e., an adequate bias vs. variance tradeoff or good achieved confidence interval coverage and width).

The adjusted coefficient of multiple determination has been used in model selection in an LS setting (the adjusted coefficient $= 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right)$, where R^2 is the usual coefficient of multiple determination; Draper and Smith 1981:91–92). Under this method, one selects the model in which this adjusted statistic is largest. McQuarrie and Tsai (1998) found this approach to be very poor (also see Rencher and Pun (1980). While adjusted R^2 is useful as a descriptive statistic, it is not useful in model selection. Mallows's C_p statistic (Mallows 1973, 1995) is also used in LS regression with normal residuals and a constant variance and in this special case provides a ranking of the candidate models that is the same as the rankings under AIC (the numerical values, C_p vs. AIC, will differ, see Atilgan 1996). The selection of models using the adjusted R^2 statistic and Mallows's C_p are related for simple LS problems (see Seber 1977:362–369). Hurvich and Tsai (1989) and McQuarrie and Tsai (1998) provide some comparisons of AIC_c vs. several competitors for linear regression problems.

Bayesian researchers have taken somewhat different approaches and assumptions, and have proposed several alternative methods for model selection. Methods such as CAIC, BIC (SIC), WIC, and HQ are mentioned in Section 2.8, as well as full Bayesian model selection (see especially Hoeting et al. 1999). These other Bayesian approaches to model selection and inference are at the current state of the art in statistics but may seem very difficult to understand and implement and are very computer intensive (e.g., Laud and Ibrahim 1995 and Carlin and Chib 1995). Draper (1995) provides a recent review of these advanced methods (also see Potscher 1991). Spiegelhalter et al. (2002) have developed a deviance information criterion (DIC) from a Bayesian perspective that is analogous to AIC. This seems to represent a blending of frequentist and Bayesian thinking, resulting in an AIC-like criterion.

The general approach that we advocate here is one derived by Akaike (1973, 1974, 1977, 1978a and b, and 1981a and b), based on information theory, and it is discussed at length in this book. Akaike's information-theoretic approach has led to a number of alternative methods having desirable properties for the selection of best approximating models in practice (e.g., AIC, AIC_c, QAIC_c, and TIC—Chapters 2 and 7). Our general advocacy concerning AIC and the associated criteria is somewhat stronger than that of Linhart and Zucchini (1986) but similar in that they also recommend objective procedures based on some well-defined criterion with a strong, fundamental basis.

1.5 Data Dredging, Overanalysis of Data, and Spurious Effects

The process of analyzing data with few or no a priori questions, by subjectively and iteratively searching the data for patterns and "significance," is often called by the derogatory term "data dredging." Other terms include "post hoc

data analysis" or "data snooping," or "data mining," but see Hand (1998) and Hand et al. (2000) for a different meaning of data mining with respect to very large data sets. Often the problem arises when data on many variables have been taken with little or no a priori motive or without benefit of supporting science. No specific objectives or alternatives were in place prior to the analysis; thus the data are submitted for analysis in the hope that the computer and a plethora of null hypothesis test results will provide information on "what is significant." A model is fit, and variables not in that model are added to create a new model, letting the data and intermediate results suggest still further models and variables to be investigated. Patterns seen in the early part of the analysis are "chased" as new variables, cross products, or powers of variables are added to the model and alternative transformations tried. These new models are clearly based on the intermediate results from earlier waves of analyses. The final model is the result of effective dredging, and often nearly everything remaining is "significant." Under this view, Hosmer and Lemeshow (1989:169) comment that "Model fitting is an iterative procedure. We rarely obtain the final model on the first pass through the data." However, we believe that such a final model is probably overfitted and unstable (i.e., likely to vary considerably if other sample data were available on the same process) with actual predictive performance (i.e., on new data) often well below what might be expected from the statistics provided by the terminal analysis (e.g., Chatfield 1996, Wang 1993). The inferential properties of a priori versus post hoc data analysis are very different. For example, (traditionally) no valid estimates of precision can be made from the model following data dredging (but see Ye 1998).

1.5.1 Overanalysis of Data

If data dredging is done, the resulting model is very much tailored (i.e., overfitted) to the data in a post hoc fashion, and the estimates of precision are likely to be overestimated. Such tailoring overdescribes the data and diminishes the validity of inferences made about the information in the data to the population of interest. Many naive applications of classical multivariate analyses are merely "fishing trips" hoping to find "significant" linear relationships among the many variables subjected to analysis (Rexstad et al. 1988, 1990, Cox and Reid 2000).

Computer routines (e.g., SAS INSIGHT) and associated manuals make data dredging both easy and "effective." Some statistical literature deals with the so-called *iterative process of model building* (e.g., Henderson and Velleman 1981). One looks for patterns in the residuals, employs various tests for selecting variables in their decreasing order of "importance," and tries all possible models. Stepwise regression and discriminant functions, for example, are used to search for "significant" variables; such methods are especially problematic if many variables (Freedman's paradox) are available for analysis (sometimes data are available on over 100 variables, and the sample size may often be less

than the number of variables). These problems of overfitting can escalate when flexible generalized linear or generalized additive models are employed.

White (2000:1097) notes, "It is widely acknowledged by empirical researchers that data snooping [dredging] is a dangerous practice to be avoided, but in fact it is endemic." Examples of data dredging include the examination of crossplots or a correlation matrix of the explanatory variables versus the response variable. These data-dependent activities can suggest apparent linear or nonlinear relationships and interactions *in the sample* and therefore lead the investigator to consider additional models. These activities should be avoided, because they probably lead to overfitted models with spurious parameter estimates and inclusion of unimportant variables as regards the *population* (Anderson et al. 2001b). The sample may be well fit, but the goal is to make a valid inference from the sample to the population. This type of data-dependent, exploratory data analysis has a place in the earliest stages of investigating a biological relationship but should probably remain unpublished. However, such cases are not the subject of this book, and we can only recommend that the results of such procedures be treated as possible hypotheses (Lindsey 1999c, Longford and Nelder 1999). New data should be collected to address these hypotheses effectively and then submitted for a comprehensive and largely a priori strategy of analysis such as we advocate here.

Two types of data dredging might be distinguished. The first is that described above; a highly interactive, data dependent, iterative post hoc approach. The second is also common and also leads to likely overfitting and the finding of effects that are actually spurious. In this type, the investigator also has little a priori information; thus "all possible models" are considered as candidates (e.g., SAS PROC REG allows this as an option). Note that the "all possible models" approach usually does not include interaction terms (e.g., $x_2 * x_5$) or various transformations such as $(x_1)^2$ or $1/x_3$ or $\log(x_2)$. In even moderate-sized problems, the number of candidate models in this approach can be very large (e.g., 20 variables $>$ a million models, 30 variables $>$ a billion models). At least this second type is not explicitly data dependent, but it is implicitly data dependent and leads to the same "sins." Also, it is usually a one-pass strategy, rather than taking the results of one set of analyses and inputting some of these into the consideration of new models. Still, in some applications, computer software often can systematically search all such models nearly automatically, and thus the strategy of trying all possible models (or at least a very large number of models) continues, unfortunately, to be popular. We believe that many situations could be substantially improved if the researcher tried harder to focus on the science of the situation before proceeding with such an unthoughtful approach.

Standard inferential tests and estimates of precision (e.g., ML or LS estimators of the sampling covariance matrix, given a model) are invalid when a final model results from the first type of data dredging. Resulting "P-values" are misleading, and there is no valid basis to claim "significance." Even conceptually there is no way to estimate precision because of the subjectivity involved

in iterative data dredging and the high probability of overfitting. In the second type of data dredging one might consider Bonferroni adjustments of the α-levels or P-values. However, if there were 1,000 models, then the α-level would be 0.00005, instead of the usual 0.05! Problems with data dredging are often linked with the problems with hypothesis testing (Johnson 1999, Anderson et al. 2000). This approach is hardly satisfactory; thus analysts have ignored the issue and merely pretended that data dredging is without peril and that the usual inferential methods somehow still apply. **Journal editors and referees rarely seem to show concern for the validity of results and conclusions where substantial data dredging has occurred. Thus, the entire methodology based on data dredging has been allowed to be perpetuated in an unthinking manner.**

We certainly encourage people to understand their data and attempt to answer the scientific questions of interest. We advocate some examination of the data prior to the formal analysis to detect obvious outliers and outright errors (e.g., determine a preliminary truncation point or the need for grouping in the analysis of distance sampling data). One might examine the residuals from a carefully chosen global model to determine likely error distributions in the candidate models (e.g., normal, lognormal, Poisson). However, if a particular pattern is noticed while examining the residuals and this leads to including another variable, then we might suggest caution concerning data dredging. Often, there can be a fine line between a largely a priori approach and some degree of data dredging.

Thus, this book will address primarily cases where there is substantial a priori knowledge concerning the issue at hand and where a relatively small set of good candidate models can be specified in advance of actual data analysis. Of course, there is some latitude where some (few) additional models might be investigated as the analysis proceeds; however, results from these explorations should be kept clearly separate from the purely a priori science. We believe that objective science is best served using a priori considerations with very limited peeking at plots of the data, parameter estimates from particular models, correlation matrices, or test statistics as the analysis proceeds. We do not condone data dredging in confirmatory analyses, but allow substantial latitude in more preliminary explorations. If some limited data dredging is done after a careful analysis based on prior considerations, then we believe that these two types of results should be carefully explained in resulting publications (Tukey 1980). For this philosophy to succeed, there should be more careful a priori consideration of alternative candidate models than has been the case in the past.

1.5.2 Some Trends

At the present time, nearly every analysis is done using a computer; thus biologists and researchers in other disciplines are increasingly using likelihood methods for more generalized analyses. Standard computer software packages

Data Dredging

Data dredging (also called data snooping, data mining, post hoc data analysis) should generally be avoided, except in (1) the early stages of exploratory work or (2) *after* a more confirmatory analysis has been done. In this latter case, the investigator should fully admit to the process that led to the post hoc results and should treat them much more cautiously than those found under the initial, a priori, approach. When done carefully, we encourage people to explore their data beyond the important a priori phase.

We recommend a substantial, deliberate effort to get the a priori thinking and models in place and try to obtain more confirmatory results; *then* explore the post hoc issues that often arise after one has seen the more confirmatory results.

Data dredging activities form a continuum, ranging from fairly trivial (venial) to the grievous (mortal). There is often a fine line between dredging and not; our advice is to stay well toward the a priori end of the continuum and thus achieve a more confirmatory result.

One can always do post hoc analyses after the a priori analysis; but one can never go from post hoc to a priori. Why not keep one's options open in this regard?

Grievous data dredging is endemic in the applied literature and still frequently taught or implied in statistics courses without the needed caveats concerning the attendant inferential problems.

Running all possible models is a thoughtless approach and runs the high risk of finding effects that are, in fact, spurious if only a single model is chosen for inference. If prediction is the objective, model averaging is useful, and estimates of precision should include model selection uncertainty. Even in this case, surely one can often rule out many models on a priori grounds.

allow likelihood methods to be used where LS methods have been used in the past. LS methods will see decreasing use, and likelihood methods will see increasing use as we proceed into the twenty-first century. Likelihood methods allow a much more general framework for addressing statistical issues (e.g., a choice of link functions and error distributions as in log linear and logistic regression models). Another advantage in a likelihood approach is that confidence intervals with good properties can be set using profile likelihood intervals. Edwards (1976), Berger and Wolpert (1984), Azzalini (1996), Royall (1997), and Morgan (2000) provide additional insights into likelihood methods, while Box (1978) provides the historical setting relating to Fisher's general methods.

During the past twenty years, modern statistical science has been moving away from traditional formal methodologies based on statistical hypothesis testing (Clayton et al. 1986, Jones and Matloff 1986, Yoccoz 1991, Bozdogan 1994, Johnson 1995, Stewart-Oaten 1995, Nester 1996, Johnson 1999, Anderson et al. 2000). The historic emphasis on hypothesis testing will

continue to diminish in the years ahead (e.g., see Quinn and Dunham 1983, Bozdogan 1994), with increasing emphasis on estimation of effects or effect sizes and associated confidence intervals (Graybill and Iyer 1994:35, Cox and Reid 2000).

Most researchers recognize that we do not conduct experiments merely to reject null hypotheses or claim statistical significance; we want deeper insights than this. We typically want to compare meaningful (i.e., plausible) alternatives, or seek information about effects and their size and precision, or are interested in causation. **There has been too much formalism, tradition, and confusion that leads people to think that statistics and statistical science is mostly about testing uninteresting or trivial null hypotheses, whereas science is much more than this. We must move beyond the traditional testing-based thinking because it is so uninformative.**

In particular, hypothesis testing for model selection is often poor (Akaike 1981a) and will surely diminish in the years ahead. There is no statistical theory that supports the notion that hypothesis testing with a fixed α level is a basis for model selection. There are not even general formal rules (or even guidelines) that rigorously define how the various P-values might be used to arrive at a final model. How does one interpret dozens of P-values, from tests with differing power, to arrive at a good model? Only *ad hoc* rules exist in this case and generally fail to result in a final parsimonious model with good inferential properties. The multiple testing issue is problematic as is the fact that likelihood ratio tests exist only for nested models. Tests of hypotheses within a data set are not independent, making inferences difficult. The order of testing is arbitrary, and differing test order will often lead to different final models. Model selection is dependent on the arbitrary choice of α, but α should depend on both n and K to be useful in model selection; however, theory for this is lacking. Testing theory is problematic when nuisance parameters occur in the models being considered. Finally, there is the fact that the so-called null is probably false on simple a priori grounds (e.g., H_0: the treatment had *no* effect, so the parameter θ is constant across treatment groups or years, $\theta_1 = \theta_2 = \cdots = \theta_k$). Rejection of such null hypotheses does not mean that the effect or parameter should be included in the approximating model! The entire testing approach is both common and somewhat absurd. All of these problems have been well known in the literature for many years; they have merely been ignored in the practical analysis of empirical data. Nester (1996) provides an interesting summary of quotations regarding hypothesis testing.

Unfortunately, it has become common to compute estimated test power after a hypothesis test has been conducted and found to be nonsignificant. Such post hoc power is not valid (Goodman and Berlin 1994, Gerard et al. 1998, Hoenig and Heisey 2001). While a priori power and sample size considerations are important in planning an experiment or observational study, estimates of post hoc power are not valid and should not be reported (Anderson et al. 2001d).

Computational restrictions prevented biologists from evaluating alternative models until the past two decades or so. Thus, people tended to use an available

model, often without careful consideration of alternatives. Present computer hardware and software make it possible to consider a number of alternative models as an integral component of data analysis. Computing power has permitted more computer-intensive methods such as the various cross-validation and bootstrapping approaches and other resampling schemes (Mooney and Duval 1993, Efron and Tibshirani 1993), and such techniques will see ever increasing use in the future.

The size or dimension (K) of some biological models can be quite high, and this has tended to increase over the past two decades. Open capture–recapture and band recovery models commonly have 20–40 estimable parameters for a single data set and might have well over 200 parameters for the joint analysis of several data sets (see Burnham et al. 1987, Preface, for a striking example of these trends). Analysis methods for structural equations commonly involve 10–30 parameters (Bollen and Long 1993). These are applications where objective model specification and selection is essential to answer the question, "*What inferences do the data support about the population*?"

1.6 Model Selection Bias

The literature on model selection methods has increased substantially in the past 15–25 years; much of this has been the result of Akaike's influential papers in the mid-1970s. However, relatively little appears in the literature concerning the properties of the parameter estimators, given that a data-dependent model selection procedure has been used (see Rencher and Pun 1980, Hurvich and Tsai 1990, Miller 1990, Goutis and Casella 1995, Ye 1998). Here, data are used to both select a parsimonious model and estimate the model parameters and their precision (i.e., the conditional sampling covariance matrix, given the selected model). These issues prompt a concern for both model selection bias and model selection uncertainty (Section 1.7).

Bias in estimates of model parameters often arises when data-based selection has been done. Miller (1990) provides a technical discussion of model selection bias in the context of linear regression. He notes his experience in the stepwise analysis of meteorological data with large sample sizes and 150 candidate models. When selecting only about 5 variables from the 150 he observed, he found t statistics as large as 6, suggesting that a particular variable was very highly significant, and yet even the sign of the corresponding regression coefficient could be incorrect. Miller warns that P-values from subset selection software are totally without foundation, and large biases in regression coefficients are often caused by data-based model selection.

Consider a linear model where there is a response variable (y) and 4 explanatory variables x_j, where $j = 1, \ldots, 4$. Order is not important in this example, so for convenience let x_1 be, in fact, very important, x_2 important, x_3 somewhat important, while x_4 is barely important. Given a decent sample size,

nearly any model selection method will indicate that x_1 and probably x_2 are important (Miller called such variables "dominant"). If one had 1,000 replicate data sets of the same size, from the same stochastic process, x_1 (particularly) and x_2 would be included in the model in nearly all cases. In these cases, an inference from a sample data set to the population would be valid. For models selected that included predictors x_1 and x_2 (essentially all 1,000 models), the estimators of the regression coefficients associated with variables x_1 and x_2 would have good statistical properties with respect to bias and precision (i.e., standard theory tends to hold for the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$).

Variable x_3 is somewhat marginal in its importance; assume, for example, that $|\beta_3|/\text{se}(\beta_3) \approx 1$, and thus its importance is somewhat small. This variable might be included in the model in only 15–30% of the 1,000 data sets. In data sets where it is selected, it tends to have an estimated regression coefficient that is biased away from zero. Thus, an inference from one of the data sets concerning the population tend to exaggerate the importance of the variable x_3. An inference from a data set in one of the remaining 70–85% of the data sets would imply that x_3 was of no importance. Neither of these cases is satisfactory.

Variable x_4 is barely important at all (a tapering effect), and it might have $|\beta_4|/\text{se}(\beta_4) \approx \frac{1}{4}$. This variable might be included in only a few (e.g., 5–10%) of the 1,000 data sets and, when it is selected, there will likely be a large bias (away from 0) in the estimator of this regression parameter. Inference from a particular sample where this variable is included in the model would imply that the variable x_4 was much more important than is actually the case (of course, the investigator has no way to know that $\hat{\beta}_4$, when selected, might be in the upper 5–10% of its sampling distribution). Then, if one examines the usual t-test, where $t = \hat{\beta}_4/\,\widehat{\text{se}}(\hat{\beta}_4)$, the likely decision will often be that the variable x_4 is significant, and should be retained in the model. This misleading result comes from the fact that the numerator in the test is biased high, while the denominator is biased low. The analyst has no way to know that this test result is probably spurious.

When predictor variables x_3 and x_4 are included in models, the associated estimator for a σ^2 is negatively biased and precision is exaggerated. These two types of bias are called model selection bias and can often be quite serious (Miller 1990, Ye 1998). Ye (1998) warns, "... the identification of a clear structure bears little cost [i.e., including variables x_1 and x_2], whereas searching through white noise has a heavy cost [i.e., including variable x_4 in a model]." Of course, in the analysis of real data, the investigator typically does not know which (if any) variables are dominant versus those that are, in fact, of marginal importance. Model selection bias is related to the problem of overfitting, the notion of tapering effect sizes, and Freedman's (1983) paradox.

The problem of model selection bias is particularly serious when little theory is available to guide the analysis. Many exploratory studies have hundreds or even thousands of models, based on a large number of explanatory variables; very often the number of models exceeds the size of the sample. Once a final model has been (somehow) selected, the analyst is usually unaware that this

model is likely overfit, with substantially biased parameter estimates (i.e., both the estimated structural regression coefficients, which are biased away from 0 and the estimated residual variation, which is biased low). They have unknowingly extracted some of the residual variation as if it represented model structure. When sample size is large, true replication exists, and there are relatively few models, these problems may be relatively unimportant. However, often one has only a small sample size, no true replication, and many models and variables; then model selection bias is usually severe (Zucchini 2000).

If, for example, x_3 is uncorrelated with x_1, x_2, and x_4, then the distribution of $\hat{\beta}_3$ is symmetric around β_3 and bias, given that x_3 is selected, is nil (i.e., if $\beta_3 = 0$, then $\mathrm{E}(\hat{\beta}_3) = 0$). This is an interesting result, but probably uncommon in practice because predictor variables are almost always correlated. Consider the case where $\beta_3 = 0$, but x_3 is highly correlated with x_1 and $\beta_1 > 0$. If the correlation between x_1 and x_3 is high (even 0.5) and positive, then when variable x_3 is selected, it is much more likely to be when $\hat{\beta}_3 > 0$. In all samples where x_3 is selected, $\hat{\beta}_3$ tends to be positive. In cases where the correlation between x_1 and x_3 is negative, then $\hat{\beta}_3$ tends to be negative. In either case, $\hat{\sigma}^2$ is biased low. By itself, x_3 would have some predictive value, but only because of its correlation with x_1, which is actually correlated with the response variable.

If sample size is small and there are many variables and hence models, then the negative bias in $\hat{\sigma}^2$ is often severe. If the predictor variables are highly intercorrelated and only one (say x_{11}) is actually correlated with the response variable, then the estimates of the regression coefficients will likely be substantially biased away from 0 in the subset of models where the associated predictor variable is selected. Leamer (1978), Copas (1983), Lehmann (1983) Gilchrist (1984), Breiman (1992), Zhang (1992*a*), and Chatfield (1995*b*, 1996) give insights into problems that arise when the same data are used both to select the model and to make inferences from that model.

1.7 Model Selection Uncertainty

Model selection uncertainty also arises when the data are used for both model selection and parameter estimation (Hjorth 1994:15–23). If a best model has been selected from a reasonable set of candidate models, bias in the model parameter estimators might be small for several of the more important variables, but might be substantial for variables associated with tapering effects. However, there is uncertainty as to the best model to use. From the example above, one must ask whether β_3 or β_4 should be in the model; this model uncertainty is a component of variance in the estimators.

Denote the sampling variance of an estimator $\hat{\theta}$, *given a model*, by $\mathrm{var}(\hat{\theta}|model)$. More generally, the sampling variance of $\hat{\theta}$ should have two components: (1) $\mathrm{var}(\hat{\theta}|model)$ and (2) a variance component due to not know-

ing the best approximating model to use (and, therefore, having to estimate this). Thus, if one uses a method such as AIC to select a parsimonious model, given the data, and estimates a conditional sampling variance, given the selected model. Then estimated precision will be too small because the variance component for model selection uncertainty is missing. Model selection uncertainty is the component of variance that reflects that model selection merely *estimates* which model is best, based on the single data set; a different model (in the fixed set of models considered) may be selected as best for a different replicate data set arising from the same experiment.

Failure to allow for model selection uncertainty often results in estimated sampling variances and covariances that are too low, and thus the achieved confidence interval coverage will be below the nominal value. Optimal methods for coping with model selection uncertainty are at the forefront of statistical research; better methods might be expected in the coming years, especially with the continued increases in computing power. Model selection uncertainty is problematic in making statistical inferences; if the goal is only data description, then perhaps selection uncertainty is a minor issue.

One must keep in mind that there is often considerable uncertainty in the selection of a particular model as the "best" approximating model. The observed data are conceptualized as random variables; their values would be different if another, independent sample were available. It is this "sampling variability" that results in uncertain statistical inference from the particular data set being analyzed. While we would like to make inferences that would be robust to other (hypothetical) data sets, our ability to do so is still quite limited, even with procedures such as AIC, with its cross-validation properties, and with independent and identically distributed sample data. Various computer-intensive resampling methods will further improve our assessment of the uncertainty of our inferences, but it remains important to understand that proper model selection is accompanied by a substantial amount of uncertainty. The bootstrap technique can effectively allow insights into model uncertainty; this and other similar issues are the subject of Chapter 5.

Perhaps we cannot totally overcome problems in estimating precision, following a data-dependent selection method such as AIC (e.g., see Dijkstra 1988, Ye 1998). This limitation certainly warrants exploration because model selection uncertainty is a quite difficult area of statistical inference. However, we must also consider the "cost" of *not* selecting a good parsimonious model for the analysis of a particular data set. That is, a model is just somehow "picked" independent of the data and used to approximate the data as a basis for inference. This procedure simply ignores both the uncertainty associated with model selection and the benefits of selection of a model that is parsimonious. This naive strategy certainly will incur substantial costs in terms of reliable inferences because model selection uncertainty is ignored (assumed to be zero). Alternatively, one might be tempted into an iterative, highly interactive strategy of data analysis (unadulterated data dredging). Again, there are substantial costs in terms of reliable inference using this approach. In particular, it seems

impossible to objectively and validly estimate the precision of the estimators following data dredging.

1.8 Summary

Truth in the biological sciences and medicine is extremely complicated, and we cannot hope to find exact truth or full reality from the analysis of a finite amount of data. Thus, inference about truth must be based on a good approximating model. Likelihood and least squares methods provide a rigorous inference theory if the model structure is "given." However, in practical scientific problems, the model is *not* "given." Thus, the critical issue is, "what is the best model to use." This is the model selection problem.

The emphasis then shifts to the careful a priori definition of a set of candidate models. This is where the science of the problem enters the analysis. Ideally, there should be a good rationale for including each particular model in the set, as well as a careful justification for why other models were excluded. The degree to which these steps can be implemented suggests a more confirmatory analysis, rather than a more exploratory analysis. Critical thinking about the scientific question and modeling alternatives, prior to looking at the data, have been underemphasized in many statistics classes in the past. These are important issues, and one must be careful not to engage in data dredging, because this weakens inferences that might be made. Information-theoretic methods provide a simple way to select a best approximating model from the candidate set of models.

In general, the information-theoretic approach should not mean merely searching for a single best model as a basis for inference. Even if model selection uncertainty is included in estimates of precision, this is a poor approach in many cases. Instead, multimodel inference should be the usual approach to making valid inference. Here, models are ranked and scaled to enhance an understanding of model uncertainty over the set. These methods are easy to understand and compute. Specific methodologies for this more general approach are the subject of this book.

We cannot overstate the importance of the scientific issues, the careful formulation of multiple working hypotheses, and the building of a small set of models to clearly and uniquely represent these hypotheses. The methods to be presented in the following chapters are "easy" to understand, compute, and interpret; however, they rest on both good science and good data that relate to the issue. We try to emphasize a more confirmatory endeavor in the applied sciences, rather than exploratory work that has become so common and has often led to so little (Anderson et al. 2000).

Data analysis is taken to mean the entire integrated process of a priori model specification, model selection, and estimation of parameters and their precision. Scientific inference is based on this process. Information-

theoretic methods free the analyst from the limiting concept that the proper approximating model is somehow "given."

The principle of parsimony is fundamental in the sciences. However, data-based selection of a parsimonious model is challenging. There are substantial rewards for proper model selection in terms of valid inferences; there are substantial dangers in either underfitting or overfitting. However, even if one has selected a good approximating model, there are issues of model selection bias and model selection uncertainty. Perhaps these cannot be fully overcome, but their effects can be lessened. These issues will be addressed in the material to follow.

Zhang (1994) notes that for the analyst who is less concerned with theoretical optimality it is more important to have available methods that are simple but flexible enough to be used in a variety of practical situations. The information-theoretic methods fall in this broad class and, when used properly, promote reliable inference.