2

Information and Likelihood Theory: A Basis for Model Selection and Inference

Full reality cannot be included in a model; thus we seek a good model to approximate the effects or factors supported by the empirical data. The selection of an appropriate approximating model is critical to statistical inference from many types of empirical data. This chapter introduces concepts from information theory (see Guiasu 1977), which has been a discipline only since the mid-1940s and covers a variety of theories and methods that are fundamental to many of the sciences (see Cover and Thomas 1991 for an exciting overview; Figure 2.1 is produced from their book and shows their view of the relationship of information theory to several other fields). In particular, the Kullback–Leibler "distance," or "information," between two models (Kullback and Leibler 1951) is introduced, discussed, and linked to Boltzmann's entropy in this chapter. Akaike (1973) found a simple relationship between the Kullback–Leibler distance and Fisher's maximized log-likelihood function (see deLeeuw 1992 for a brief review). This relationship leads to a simple, effective, and very general methodology for selecting a parsimonious model for the analysis of empirical data.

Akaike introduced his "*entropy maximization principle*" in a series of papers in the mid-1970s (Akaike 1973, 1974, 1977) as a theoretical basis for model selection. He followed this pivotal discovery with several related contributions beginning in the early 1980s (Akaike 1981a and b, 1985, 1992, and 1994). This chapter introduces AIC and related criteria such as AIC_c, QAIC_C, and TIC. No mathematical derivations of these criteria are given here because they are given in full detail in Chapter 7. We urge readers to understand the full derivation (given in Chapter 7), for without it, the simple and compelling idea

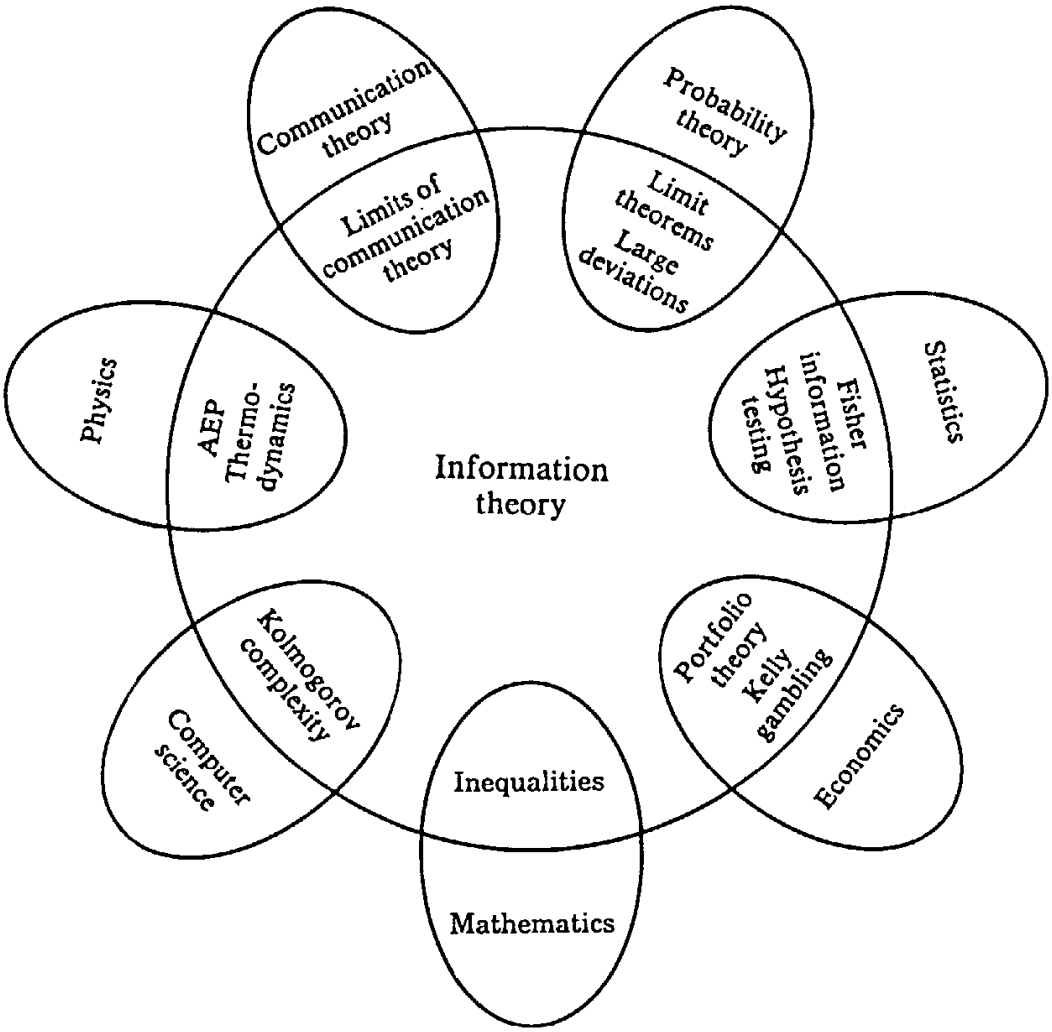

FIGURE 2.1. Information theory and its relationships to other disciplines (from Cover and Thomas 1991). Information theory began in the mid-1940s, at the close of WWII. In the context of this book, the most relevant components of information theory include Fisher information, entropy (from thermodynamics and communication theory), and Kullback–Leibler information.

underlying Kullback–Leibler information and the various information criteria cannot be fully appreciated.

2.1 Kullback–Leibler Information or Distance Between Two Models

We begin without any issues of parameter estimation and deal with very simple expressions for the models f and g, assuming that they are completely known. In initial sections of this chapter we will let both f and g be simple probability distributions, since this will allow an understanding of K-L information or distance in a simple setting. However, we will soon switch to the concept that

f is a notation for full reality or truth. We use g to denote an approximating model in terms of a probability distribution.

Kullback–Leibler Information

Kullback-Leibler information between models f and g is defined for continuous functions as the (usually multi-dimensional) integral

$$\boldsymbol{I(f,g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx,}$$

where log denotes the natural logarithm. The notation $I(f, g)$ denotes the **"information lost when g is used to approximate f."**

As a heuristic interpretation, **$I(f,g)$ is the distance from g to f**.

We will use both interpretations throughout this book, since both seem useful. Of course, we seek an approximating model that loses as little information as possible; this is equivalent to minimizing $I(f, g)$, over g. Full reality f is considered to be given (fixed), and only g varies over a space of models indexed by θ. Similarly, Cover and Thomas (1991) note that the K-L distance is a measure of the inefficiency of assuming that the distribution is g when the true distribution is f.

Kullback–Leibler Information

The expression for the Kullback-Leibler information or distance in the case of discrete distributions such as the Poisson, binomial, or multinomial is

$$\boldsymbol{I(f,g) = \sum_{i=1}^{k} p_i \cdot \log \left(\frac{p_i}{\pi_i} \right).}$$

Here, there are k possible outcomes of the underlying random variable; the true probability of the ith outcome is given by p_i, while the $\pi_1, \ldots, \pi_k$ constitute the approximating probability distribution (i.e., the approximating model). In the discrete case, we have $0 < p_i < 1$, $0 < \pi_i < 1$, and $\sum p_i = \sum \pi_i = 1$. Hence, here f and g correspond to the p_i and π_i, respectively.

As in the continuous care the notation **$I(f,g)$** denotes the **information lost when g is used to approximate f or the distance from g to f**.

In the following material we will generally think of K-L information in the continuous case and use the notation f and g for simplicity.

Well over a century ago measures were derived for assessing the "distance" between two models or probability distributions. Most relevant here is Boltzmann's (1877) concept of generalized entropy (see Section 2.12) in physics and thermodynamics (see Akaike 1985 for a brief review). Shannon (1948) employed entropy in his famous treatise on communication theory (see Atmar 2001 for an exciting review of information theory, its practicality, and relations to evolution). Kullback and Leibler (1951) derived an information measure that

Ludwig Eduard Boltzmann, 1844–1906, one of the most famous scientists of his time, made incredible contributions in theoretical physics. He received his doctorate in 1866; most of his work was done in Austria, but he spent some years in Germany. He became full professor of mathematical physics at the University of Graz, Austria, at the age of 25. His mathematical expression for entropy was of fundamental importance throughout many areas of science. The negative of Boltzmann's entropy is a measure of "information" derived over half a century later by Kullback and Leibler. J. Bronowski wrote that Boltzmann was "an irascible, extraordinary man, an early follower of Darwin, quarrelsome and delightful, and everything that a human should be." Several books chronicle the life of this great figure of science, including Cohen and Thirring (1973) and Broda (1983); his collected technical papers appear in Hasenöhrl (1909).

happened to be the negative of Boltzmann's entropy, now referred to as the Kullback–Leibler (K-L) information or distance (but see Kullback 1987, where he preferred the term *discrimination information*). The motivation for Kullback and Leibler's work was to provide a rigorous definition of "information" in relation to Fisher's "sufficient statistics." The K-L distance has also been called the K-L discrepancy, divergence, information, and number. We will treat these terms as synonyms, but tend to use *distance* or *information* in the material to follow.

The Kullback–Leibler distance can be conceptualized as a directed "distance" between two models, say f and g (Kullback 1959). Strictly speaking, this is a measure of "discrepancy"; it is not a simple distance, because the measure from f to g is not the same as the measure from g to f; it is a directed, or oriented, distance (Figure 2.2). The K-L distance is perhaps the

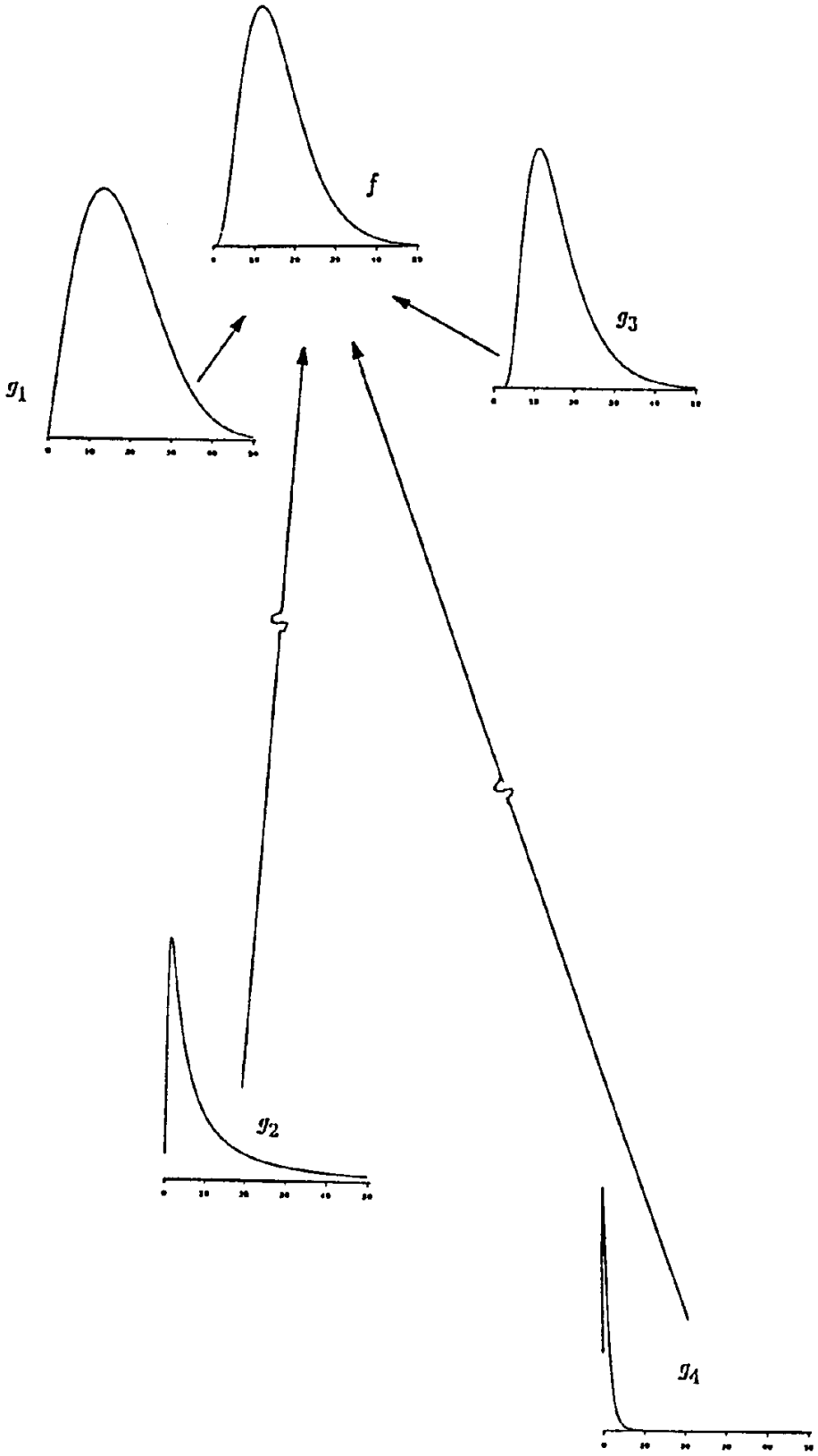

FIGURE 2.2. The Kullback–Leibler discrepancy $I(f, g_i)$ is a directed distance from the various candidate models g_i to f. Knowing the K-L distances would allow one to find which of the 4 approximating models is *closest* to model f. Here, f is gamma (4, 4), and the 4 approximating models are g_1 = Weibull (2, 20), g_2 = lognormal (2, 2), g_3 = inverse Gaussian (16, 64), and g_4 = F distribution (4, 10). In each case, the model parameters are known exactly (not estimated).

most fundamental of all information measures in the sense of being derived from minimal assumptions and its additivity property. The K-L distance is an extension of Shannon's concept of information (Hobson and Cheng 1973, Soofi 1994) and is sometimes called a "relative entropy." The K-L distance between models is a *fundamental quantity* in science and information theory (see Akaike 1983) and is the logical basis for model selection in conjunction with likelihood inference.

At a heuristic level, "information" is defined as $-\log_e(f(x))$ for some continuous probability density function or $-\log_e(p_i)$ for the discrete case. Kullback–Leibler information is a type of "cross entropy," a further generalization. In either the continuous or discrete representation, the right-hand side is an expected value (i.e., $\int f(x)\big(\cdot\big)dx$ for the continuous case or $\sum_{i=1}^{k} p_i\big(\cdot\big)$ for the discrete case) of the logarithm of the ratio of the two distributions (f and g) or two discrete probabilities (p_i and π_i). In the continuous case one can think of this as an average (with respect to f) of $\log_e(f/g)$, and in the discrete case it is an average (with respect to the p_i) of the logarithm of the ratio (p_i/π_i). The foundations of these expressions are both deep and fundamental (see Boltzmann 1877, Kullback and Leibler 1951, or contemporary books on information theory).

The K-L distance ($I(f, g)$) is always positive, except when the two distributions f and g are identical (i.e., $I(f, g) = 0$ if and only if $f(x) = g(x)$ everywhere). More detail and extended notation will be introduced in Chapter 7; here we will employ a simple notation and use it to imply considerable generality in the sample data (x) and the multivariate functions f and g.

2.1.1 Examples of Kullback–Leibler Distance

An example will illustrate the K-L distances ($I(f, g_i)$). Let f be a gamma distribution with 2 parameters ($\alpha = 4$, $\beta = 4$). Then consider 4 approximating models g_i, each with 2 parameters (see below): Weibull, lognormal, inverse Gaussian, and the F distribution. Details on these simple probability models can be found in Johnson and Kotz (1970). The particular parameter values used for the four g_i are not material here, except to stress that they are assumed known, not estimated. "Which of these parametrized distributions is the *closest* to f?" is answered by computing the K-L distance between each g_i and f (Figure 2.2). These are as follows:

	Approximating model	$I(f, g_i)$	Rank
g_1	Weibull distribution ($\alpha = 2$, $\beta = 20$)	0.04620	1
g_2	lognormal distribution ($\theta = 2$, $\sigma^2 = 2$)	0.67235	3
g_3	inverse Gaussian ($\alpha = 16$, $\beta = 64$)	0.06008	2
g_4	F distribution ($\alpha = 4$, $\beta = 10$)	5.74555	4

Here, the Weibull distribution is closest to (loses the least information about) f, followed by the inverse Gaussian. The lognormal distribution is a poor third,

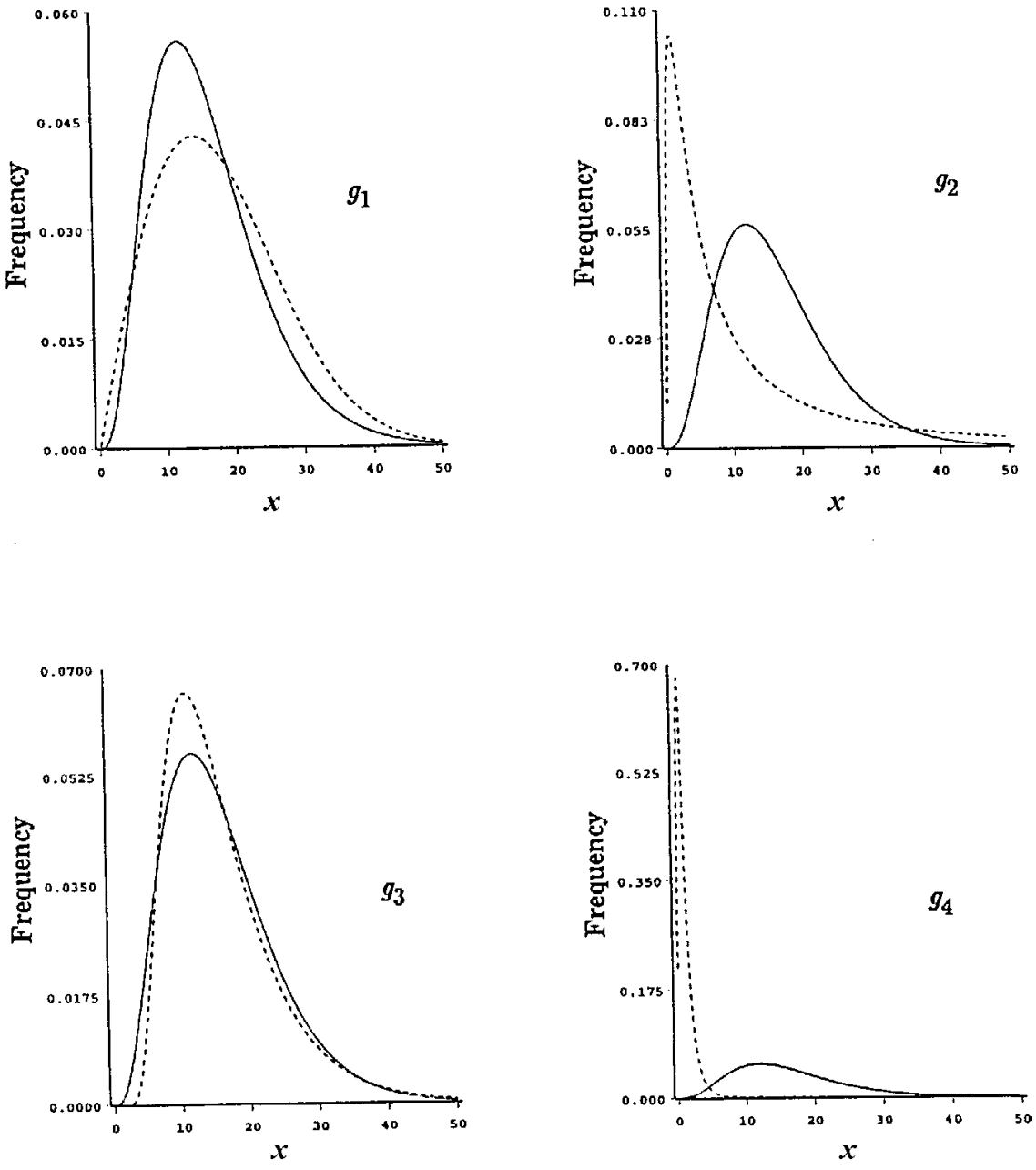

FIGURE 2.3. Plots of f (= gamma (4, 4), solid line) against each of the 4 approximating models g_i (dashed lines) as a function of x. Here, g_1 = Weibull (2, 20), g_2 = lognormal (2, 2), g_3 = inverse Gaussian (16, 64), and g_4 = F distribution (4, 10). Only in the simplest cases can plots such as these be used to judge closeness between models. Model f is the same in all 4 graphs; it is merely scaled differently to allow the $g_i(x)$ to be plotted on the same graph.

while the F distribution is relatively far from the gamma distribution f (see Figure 2.3).

Further utility of the K-L distance can be illustrated by asking which of the approximating models g_i might be closest to f when the parameters of g_i are allowed to vary (i.e., what parameter values make each g_i optimally close to f?). Following a computer search of the parameter space for the Weibull, we found that the *best* Weibull had parameters $\alpha = 2.120$ and $\beta = 18.112$ and a

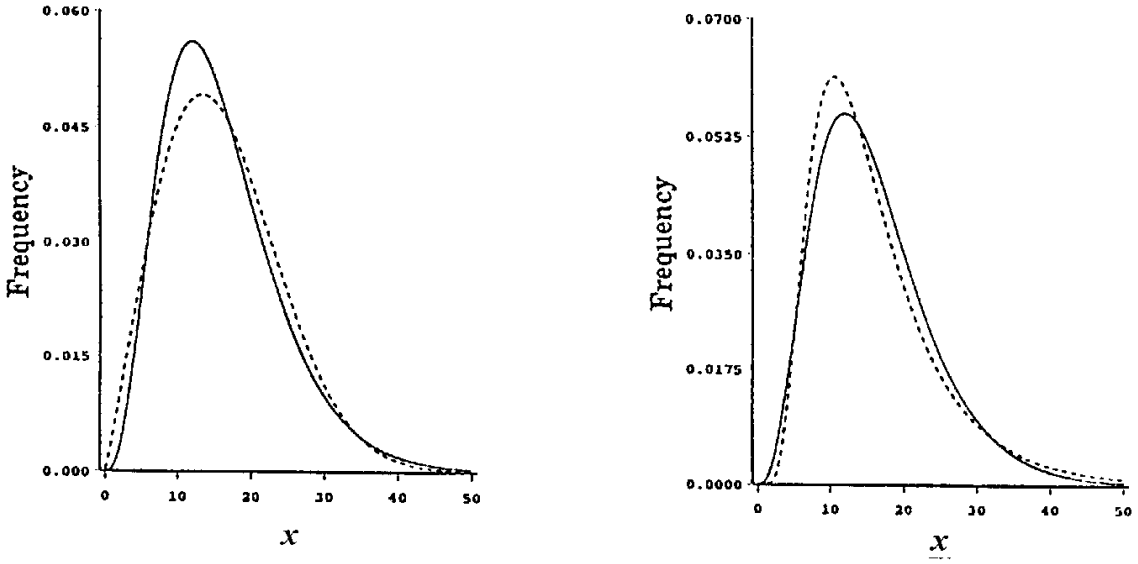

FIGURE 2.4. Plots of f (= gamma (4, 4)) against the best Weibull (left) and lognormal models. The Weibull model that was closest to f had parameters (2.120, 18.112) with K-L distance = 0.02009, while the best lognormal had parameters (2.642, 0.2838) with K-L distance = 0.02195. Compare these optimally parametrized models with those in Figure 2.3 (top).

K-L distance of 0.02009; this is somewhat closer than the original parametrization 0.04620 above. Using the same approach, the best lognormal model had parameters $\theta = 2.642$ and $\sigma^2 = 0.2838$ and a K-L distance of 0.02195, while the best inverse Gaussian model had parameters $\alpha = 16$ and $\beta = 48$ with a K-L distance of 0.03726, and the approximately best F distribution had parameters $\alpha \approx 300$, $\beta = 0.767$ and a K-L distance of approximately 1.486 (the K-L distance is not sensitive to α in this case, but is quite difficult to evaluate numerically). Thus, K-L distance indicates that the best Weibull is closer to f than is the best lognormal (Figure 2.4). Note that the formal calculation of K-L distance requires knowing the true distribution f as well as all the parameters in the models g_i (i.e., parameter estimation has not yet been addressed). Thus, K-L distance cannot be computed for real-world problems.

These values represent *directed* distances; in the first Weibull example, $I(f, g_1) = 0.04620$, while $I(g_1, f) = 0.05552$ (in fact, we would rarely be interested in $I(g_1, f)$ since this is the information lost when f is used to approximate g!). The point here is that these are directed or oriented distances and $I(f, g_1) \neq I(g_1, f)$; *nor should they be equal, because the roles of truth and model are not interchangeable.*

These are all univariate functions; thus one could merely plot them on the same scale and visually compare each g_i to f; however, this graphical method will work only in the simplest cases. In addition, if two approximating distributions are fairly close to f, it might be difficult to decide which is better by only visual inspection. Values of the K-L distance are not based on only the mean and variance of the distributions; rather, the distributions in their entirety are the subject of comparison.

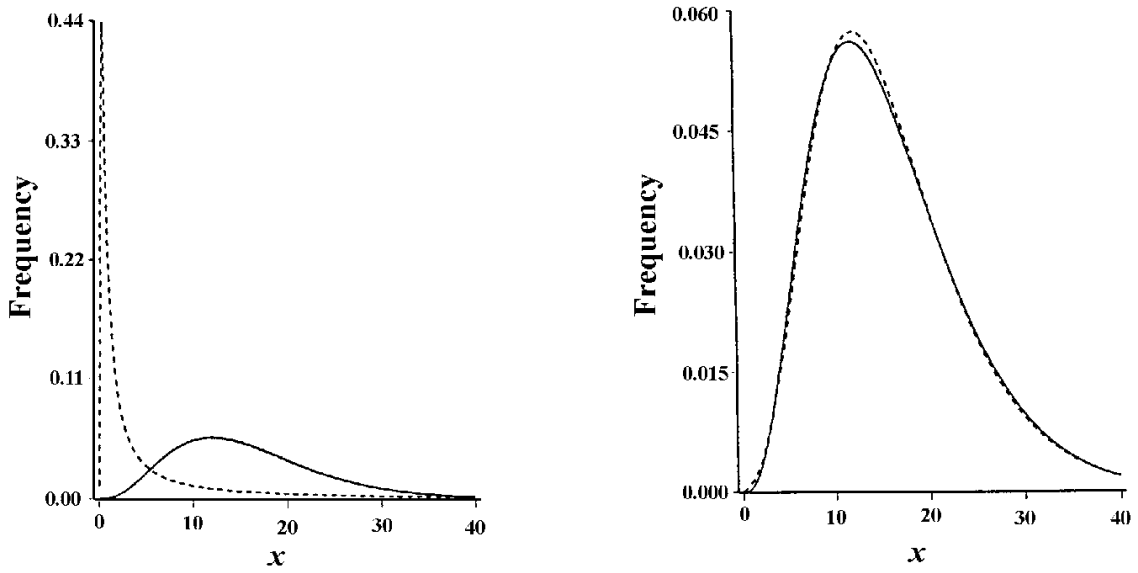

FIGURE 2.5. Plots of f (= gamma (4, 4)) against the best 2-parameter F distribution (left) and the best 3-parameter (noncentral) F distribution. The best 2-parameter model was a poor approximation to f (K-L distance = 1.486), while the best 3-parameter model is an excellent approximation (parameters 1.322, 43.308, 18.856) with K-L distance = 0.001097. Approximating models with increasing numbers of parameters typically are closer to f than approximating models with fewer parameters.

The F distribution ($\alpha = 4$, $\beta = 10$) provided a relatively poor approximation to the gamma distribution with ($\alpha = 4$, $\beta = 4$). Even the best 2-parameter F distribution remains a relatively poor approximation (K-L distance = 1.486). However, in general, adding more parameters will result in a closer approximation (e.g., the classic use of the Fourier series in the physical sciences or Wel's (1975) elephant-fitting problem). If we allow the addition of a third parameter (λ) in the F distribution (the noncentral F distribution), we find that the best model ($\alpha = 1.322$, $\beta = 43.308$, and $\lambda = 18.856$) has a K-L distance of only 0.001097; this is better than any of the other 2-parameter candidate models (Figure 2.5). Closeness of approximation can always be increased by adding more parameters to the candidate model. When we consider *estimation* of parameters and the associated uncertainty, then the principle of parsimony must be addressed (see Section 1.4), or overfitted models will be problematic.

In the remainder of the book we will want a more general, conceptual view of f, and we will use it to reflect truth or full reality. Here, reality is rarely (if ever) a model; rather, it reflects the complex biological (and measuring or sampling) process that generated the observed data x. For this reason we will not explicitly parametrize the complex function f, because it represents full reality (truth), it might not even have parameters in a sense that would be analogous to θ in a modeling framework. In fact, thinking that truth is parametrized is itself a type of (artificial) model-based conceptualization. Sometimes it is useful to think of f as full reality and let it have (conceptually) an infinite number of parameters (see Section 1.2.4). This "crutch" of infinite-dimensionality at least retains the concept of reality even though it is in some unattainable perspective. Thus, f

represents full truth, and might be conceptually based on a very large number of parameters (of a type we may have not even properly conceived) that give rise to a set of data x. Finally, we will see how this conceptualization of reality (f) collapses into a nonidentifiable constant in the context of model selection.

2.1.2 *Truth, f, Drops Out as a Constant*

The material above makes it obvious that both f and g (and their parameters) must be known to compute the K-L distance between these two models. However, if only relative distance is used, this requirement is diminished, since $I(f, g)$ can be written equivalently as

$$I(f, g) = \int f(x) \log(f(x)) dx - \int f(x) \log(g(x \mid \theta)) dx.$$

Note that each of the two terms on the right of the above expression is a statistical expectation with respect to f (truth). Thus, the K-L distance (above) can be expressed as a difference between two statistical expectations,

$$I(f, g) = \mathrm{E}_f \left[\log(f(x))\right] - \mathrm{E}_f[\log(g(x \mid \theta))],$$

each with respect to the distribution f. This last expression provides easy insights into the derivation of AIC.

The first expectation $\mathrm{E}_f[\log(f(x))]$ is a constant that depends only on the unknown true distribution, and it is clearly not known (i.e., we do not know f in actual data analysis). Therefore, treating this unknown term as a constant, a measure of *relative* directed distance is possible (Bozdogan 1987, Kapur and Kesavan 1992:155). Clearly, if one computed the second expectation $\mathrm{E}_f[\log(g(x \mid \theta))]$, one could estimate $I(f, g)$ up to a constant C (namely $\mathrm{E}_f[\log(f(x))]$),

$$I(f, g) = C - \mathrm{E}_f[\log(g(x \mid \theta))],$$

or

$$I(f, g) - C = -\mathrm{E}_f[\log(g(x \mid \theta))].$$

The term $\big(I(f, g) - C\big)$ is a *relative* directed distance between f and g; thus, $\mathrm{E}_f \left[\log(g(x \mid \theta))\right]$ becomes the quantity of interest for selecting a best model. For two models g_1 and g_2, if $I(f, g_1) < I(f, g_2)$, so g_1 is best, then $I(f, g_1) - C < I(f, g_2) - C$, and hence $-E_f[\log(g_1(x|\theta))] < -E_f[\log(g_2(x|\theta))]$. Moreover, $I(f, g_2) - I(f, g_1) \equiv -E_f[\log(g_2(x|\theta))] + E_f[\log(g_1(x|\theta))]$, so we know how much better model g_1 is than model g_2. Without knowing C we just do not know the absolute measure of how good even g_1 is, but we can identify the fact that model g_1 is better than g_2. Note that no parameter estimation is involved here, but the concepts carry over to the cases where estimation occurs. From the preceding example, where f is gamma (4, 4), then $\int f(x) \log(f(x)) dx = 3.40970$, and this term is constant across the models

being compared,

$$I(f, g) - 3.40970 = -\mathrm{E}_f[\log(g(x \mid \theta))].$$

The *relative* distances between the gamma $(4, 4)$ model and the four approximating models are shown below:

	Approximating model	Relative distance $I(f, g_i) - C$	Rank
g_1	Weibull distribution ($\alpha = 2, \beta = 20$)	3.45591	1
g_2	lognormal distribution ($\theta = 2, \sigma^2 = 2$)	4.08205	3
g_3	inverse Gaussian ($\alpha = 16, \beta = 64$)	3.46978	2
g_4	F distribution ($\alpha = 4, \beta = 10$)	9.15525	4

Note that the ranking of "closeness" of the four candidate models to f is preserved, and the relative ranking of distance between models remains unchanged, even though only relative distances are used.

Kullback-Leibler distance $I(f, g)$ is on a true ratio scale, where there is a true zero. In contrast, $-\int f(x)(\log(g(x|\theta)))dx \equiv -\mathrm{E}_f[\log(g(x|\theta))]$ is on an interval scale and lacks a true zero. A difference of magnitude D means the same thing anywhere on the scale. Thus, $D = 10 = 12 - 2 = 1012 - 1002$; a difference of 10 means the same thing anywhere on the interval scale. Then, $10 = V_1 - V_2$, regardless of the size of V_1 and V_2.

The calculation of the two components of K-L distance (above) is in effect based on a sample size of 1. If the sample size were 100, then each component would be 100 times larger, and the difference between the two components would also be 100 times larger. For example, if $n = 100$, then $\int f(x)\log(f(x))dx = 3.40970 \times 100 = 340.970$ and $\mathrm{E}_f[\log(g_1(x \mid \theta))]$ (the Weibull) $= 3.45591 \times 100 = 345.591$. Thus, the difference between the two components of K-L distance would be 4.620; the *relative* difference is large when sample size is large. A large sample size magnifies the separation of research hypotheses and the models used to represent them. **Adequate sample size conveys a wide variety of advantages in making valid inferences.**

Typically, as in the example above, the analyst would postulate several a priori candidate models $g_i(x \mid \theta)$ and want to select the *best* among these as a basis for data analysis and inference. Definition of "best" will involve the principle of parsimony and the related concept of a best approximating model. In data analysis, the parameters in the various candidate models are not known and must be estimated from the empirical data. This represents an important distinction from the material above, since one usually has only models with estimated parameters, denoted by $g_i(x \mid \hat{\theta})$. In this case, one needs *estimates* of the relative directed distances between the unknown f that generated the data and the various candidate models $g_i(x \mid \hat{\theta})$. Then, knowing the estimated relative distance from each $g_i(x)$ to $f(x)$, we select the candidate model that is *estimated* to be closest to truth for inference (Figure 2.2). That is, we select the model with the smallest estimated, *relative* distance. Alternatively, we select an approximating model that loses the least information about truth. The

conceptual truth f becomes a constant term, and nothing need be assumed about f, since the constant is the same across the candidate models and is irrelevant for comparison. (Similarly, it is interesting to note that often the log-likelihood function also involves an additive constant that is the same across models; this term is known, but generally ignored, since it is often difficult to compute.) In practice, we can obtain only an *estimator* of the relative K-L distance from each approximating model $g_i(x \mid \hat{\theta})$ to f.

2.2 Akaike's Information Criterion: 1973

Akaike's (1973) seminal paper proposed the use of the Kullback-Leibler information or distance as a fundamental basis for model selection. However, K-L distance cannot be computed without full knowledge of both f (full reality) and the parameters (θ) in each of the candidate models $g_i(x|\theta)$. Akaike found a rigorous way to estimate K-L information, based on the empirical log-likelihood function at its maximum point.

Given a parametric structural model there is a unique value of θ that, in fact, minimizes K-L distance $I(f, g)$. This (unknown) minimizing value of the parameter depends on truth f, the model g through its structure, the parameter space, and the sample space (i.e., the structure and nature of the data that can be collected). In this sense there is a "true" value of θ underling ML estimation, let this value be θ_0. Then θ_0 is the absolute best value of θ for model g; actual K-L information loss is minimized at θ_0. If one somehow knew that model g was, in fact, the K-L best model, then the MLE $\hat{\theta}$ would estimate θ_0. This property of the model $g(x|\theta_0)$ as the minimizer of K-L, over all $\theta \in \Theta$, is an important feature involved in the derivation of AIC (Chapter 7).

In data analysis the model parameters must be estimated, and there is usually substantial uncertainty in this estimation. Models based on estimated parameters, hence on $\hat{\theta}$ not θ, represent a major distinction from the case where model parameters would be known. This distinction affects how we must use K-L distance as a basis for model selection. The difference between having θ or θ_0 (we do not) and having the estimate $\hat{\theta}$ (we do) is quite important and basi-

Selection Target

Akalke (1973, 1974, 1985, 1994) showed that the critical issue for getting an applied K-L model selection criterion was to estimate

$$\mathbf{E}_y\mathbf{E}_x[\log(g(x|\hat{\theta}(y)))],$$

where x and y are independent random samples from the same distribution and both statistical expectations are taken with respect to truth (f). This double expectation, both with respect to truth f, is the target of all model selection approaches, based on K-L information.

cally causes us to change our model selection criterion to that of minimizing *expected* estimated K-L distance rather than minimizing known K-L distance over the set of R models considered.

It is tempting to just estimate $\mathrm{E}_y\mathrm{E}_x[\log(g(x|\hat{\theta}(y)))]$ by the maximized $\log(\mathcal{L}(\hat{\theta})|data)$ for each model g_i. However, Akalke (1973) showed that the maximized log-likelihood is biased upward as an estimator of the model selection target (above). He also found that under certain conditions (these conditions are important, but quite technical) this bias is approximately equal to K, the number of estimable parameters in the approximating model. This is an asymptotic result of fundamental importance.

The Key Result

Thus, an approximately unbiased estimator of

$$\mathbf{E}_y\mathbf{E}_x[\log(g(x|\hat{\theta}(y)))]$$

for large samples and "good" models is

$$\log(\mathcal{L}(\hat{\theta}|data)) - K.$$

This result is equivalent to

$$\log(\mathcal{L}(\hat{\theta}|data)) - K = \text{constant} - \hat{\mathbf{E}}_{\hat{\theta}}[I(f, \hat{g})],$$

where $\hat{g} = g(\cdot|\hat{\theta})$.

The bias-correction term ($K =$ the number of estimable parameters) above is a special case of a more general result derived by Takeuchi (1976) and described in the following section and in Chapter 7. **Akaike's finding of a relation between the relative expected K-L distance and the maximized log-likelihood has allowed major practical and theoretical advances in model selection and the analysis of complex data sets** (see Stone 1982, Bozdogan 1987, and deLeeuw 1992).

Akaike's Information Criterion

Akaike (1973) then defined "*an information criterion*" (AIC) by multiplying $\log(\mathcal{L}(\hat{\theta}|y)) - K$ by -2 ("taking historical reasons into account") to get

$$\mathbf{AIC} = -2\log(\mathcal{L}(\hat{\theta}|y)) + 2K.$$

This has become known as **"Akaike's information criterion"** or AIC.

Thus, rather than having a simple measure of the directed distance between two models (i.e., the K-L distance), one has instead an *estimate* of the expected, relative distance between the fitted model and the unknown true mechanism (perhaps of infinite dimension) that actually generated the observed data.

The expression $\log(\mathcal{L}(\hat{\theta}|y))$ is the numerical value of the log-likelihood at its maximum point (see Section 1.2.2). This maximum point on the log-likelihood

function corresponds to the values of the maximum likelihood estimates. The number of estimable parameters in the model is denoted by K, and it is usually clear as to what the correct count should be (see below for standard linear models). In some types of models there are some parameters that are not uniquely estimable from the data, and these should not be counted in K. Nonestimability can occur in the analysis of count data where a cell has no observations, and thus a parameter that is identifiable becomes nonestimable for that data set. Nonestimability can also arise due to inherent confounding (e.g., the parameters S_{t-1} and f_t in certain band recovery models of Brownie et al. 1985). In application, one computes AIC for each of the candidate models and selects the model with the smallest value of AIC. It is this model that is estimated to be "closest" to the unknown reality that generated the data, from among the candidate models considered. This seems a very natural, simple concept; select the fitted approximating model that is estimated, on average, to be closest to the unknown f. Basing AIC on the expectation (over $\hat{\theta}$) of $\mathrm{E}_x[\log(g(x|\hat{\theta}(y)))$ provides AIC with a cross-validation property for independent and identically distributed samples (see Stone 1977, Stoica et al. 1986, Tong 1994). Golub et al. (1979) show that AIC asymptotically coincides with generalized cross-validation in subset regression (also see review by Atilgan 1996).

Of course, models not in the set remain out of consideration. AIC is useful in selecting the best model in the set; however, if all the models are very poor, AIC will still select the one estimated to be best, but even that relatively best model might be poor in an absolute sense. Thus, every effort must be made to ensure that the set of models is well founded.

$I(f, g)$ can be made smaller by adding more known (not estimated) parameters in the approximating model g. Thus, for a fixed data set, the further addition of parameters in a model g_i will allow it to be closer to f. However, when these parameters must be estimated (rather than being known or "given"), further uncertainty is added to the *estimation* of the relative K-L distance. At some point, the addition of still more estimated parameters will have the opposite from desired effect (i.e., to reduce $\mathrm{E}_{\hat{\theta}}[I(f, \hat{g})]$ as desired). At that point, the estimate of the relative K-L distance will increase because of "noise" in estimated parameters that are not really needed to achieve a good model. This phenomenon can be seen by examination of the information criterion being minimized,

$$\mathbf{AIC} = -2\log(\mathcal{L}(\hat{\theta}|y)) + 2K,$$

where the first term on the right-hand side tends to decrease as more parameters are added to the approximating model, while the second term ($2K$) gets larger as more parameters are added to the approximating model. This is the tradeoff between bias and variance or the tradeoff between underfitting and overfitting that is fundamental to the principle of parsimony (see Section 1.4.2). Some investigators have considered K to be a measure of "complexity," but this is unnecessary, though not irrational. We consider K primarily a

simple expression for the asymptotic bias in the log-likelihood as an estimator of $E_y E_x[\log(g(x|\hat{\theta}(y)))]$. Note that AIC is derived as an estimator of relative, expected K-L information; thus parsimony arises as a byproduct of this approach. Further books and papers on the derivation of AIC include Shibata (1983, 1989), Linhart and Zucchini (1986), Bozdogan (1987), and Sakamoto (1991).

Usually, AIC is positive; however, it can be shifted by any additive constant, and some shifts can result in negative values of AIC. Computing AIC from regression statistics (see Section 1.2.2) often results in negative AIC values. In our work, we have seen minimum AIC values that range from large negative numbers to as high as 340,000. **It is not the absolute size of the AIC value, it is the relative values over the set of models considered, and particularly the differences between AIC values (Section 2.5), that are important.**

The material to this point has been based on likelihood theory, which is a very general approach. In the special case of least squares (LS) estimation with normally distributed errors, and apart from an arbitrary additive constant, AIC can be expressed as a simple function of the residual sum of squares.

The Least Squares Case

If all the models in the set assume normally distributed errors with a constant variance, then AIC can be easily computed from least squares regression statistics as

$$\mathbf{AIC} = n \log(\hat{\sigma}^2) + 2K,$$

where

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n} \text{ (the MLE of } \sigma^2),$$

and $\hat{\epsilon}_i$ are the estimated residuals for a particular candidate model. A common mistake with LS model fitting, when computing AIC, is to take the estimate of σ^2 from the computer output, instead of computing the ML estimate, above. **Also, for LS model fitting, K is the total number of estimated regression parameters, including the intercept and σ^2.**

Thus, AIC is easy to compute from the results of LS estimation in the case of linear models and is now included in the output of many software packages for regression analysis. However, the value of K is sometimes determined incorrectly because either β_0 (the intercept) or σ^2 (or both) is mistakenly ignored in determining K.

The fact that AIC is an estimate only of relative expected K-L distance is almost unimportant. It is the fact that AIC is only an estimate of these relative distances from each model g_i to f that is less than ideal. It is important to recognize that there is usually substantial uncertainty as to the best model for a given data set. After all, these are stochastic biological processes, often with relatively high levels of uncertainty.

In as much as a statistical model can provide insight into the underlying biological process, it is important to try to determine as accurately as possible the basic underlying structure of the model that fits the data well. "Let the data speak" is of interest to both biologists and statisticians in objectively learning from empirical data. The data then help determine the proper complexity (order or dimension) of the approximating model used for inference and help determine what effects or factors are justified. In this sense, inferences for a given data set are conditional on sample size. We must admit that if much more data were available, then further effects could probably be found and supported. "Truth" is elusive; model selection tells us what inferences the data support, not what full reality might be.

Akaike (1973) multiplied the bias-corrected log-likelihood by -2 for "historical reasons" (e.g., it is well known that -2 times the logarithm of the ratio of two maximized likelihood values is asymptotically chi-squared under certain conditions and assumptions). The term -2 occurs in other statistical contexts, so it was not unreasonable that Akaike performed this simple operation to get his AIC. Two points frequently arise, and we will note these here. First, the model associated with the minimum AIC remains unchanged if the bias-corrected log-likelihood (i.e., $\log(\mathcal{L}) - K$) is multiplied by -0.17, -34, or -51.3, or any other negative number. Thus, the minimization is not changed by the multiplication of both terms by any negative constant; Akaike merely chose -2. Second, some investigators have not realized the formal link between K-L information and AIC and believed, then, that the number 2 in the second term in AIC was somehow "arbitrary" and that other numbers should also be considered. This error has led to considerable confusion in the technical literature; clearly, K is the asymptotic bias correction and is not arbitrary. Akaike chose to work with $-2\log(\mathcal{L})$, rather than $\log(\mathcal{L})$; thus the term $+2K$ is theoretically correct, for large sample size. As long as *both* terms (the log-likelihood and the bias correction) are multiplied by the same negative constant, the model where the criterion is minimized is unchanged and there is nothing arbitrary.

It might be argued that we should have merely defined $l = \log(\mathcal{L}(\hat{\theta} \mid \textit{data, model}))$; then AIC $= -2l + 2K$, making the criterion look simpler. While this may have advantages, we believe that the full notation works for the reader and helps in understanding exactly what is meant. The full notation, or abbreviations such as $\log(\mathcal{L}(\theta|x, g_i))$, makes it explicit that the log-likelihood is a function of (only) the parameters (θ), while the data (x) and model (g_i, say multinomial) must be given (i.e., known). These distinctions become more important when we introduce the concept of a likelihood of a model, given the data: $\mathcal{L}(g_i|\textit{data})$. Both concepts are fundamental and useful in a host of ways in this book and the notation serves an important purpose here.

If the approximating models in the candidate set are poor (far from f), then Takeuchi's information criterion (TIC) is an alternative if sample size is quite large. AIC is a special case of TIC, and as such, AIC is a parsimonious approach to the estimation of relative expected K-L distance (see Section 2.3).

2.3 Takeuchi's Information Criterion: 1976

At one point in Akaike's derivation of an estimator of K-L information he made the assumption that the model set included f (full reality). This has been the subject of attention and criticism. Akaike maintained that his estimator (AIC) was asymptotically unbiased and free from any notion that full reality was a model or that such a true model was required to be in the set of candidate models. This section will indicate that such claims were justified and provides another insight into the concept of parsimony. The key to this issue is an important, little-known paper (in Japanese) by Takeuchi (1976) that appeared just 3 years after Akaike's initial breakthrough in 1973.

Takeuchi (1976) provides a very general derivation of an information criterion, without taking expectations with respect to g. His criterion is now called TIC (Takeuchi's information criterion) and was thought to be useful in cases where the candidate models were not particularly close approximations to f. TIC has a more general bias-adjustment term to allow $-2\log(\mathcal{L})$ to be adjusted to be an asymptotically unbiased estimate of relative, expected K-L information,

$$\mathbf{TIC = -2\log(\mathcal{L}) + 2 \cdot tr(J(\theta)I(\theta)^{-1}).}$$

The $K \times K$ matrices $J(\theta)$ and $I(\theta)$ involve first and second mixed partial derivatives of the log-likelihood function, and "tr" denotes the matrix trace function. One might consider always using TIC and worry less about the adequacy of the models in the set of candidates. This consideration involves two issues that are problematic. First, one must *always* worry about the quality of the set of approximating models being considered; this is not something to shortcut. Second, using the expanded bias adjustment term in TIC involves estimation of the elements of the matrices $J(\theta)$ and $I(\theta)$ (details provided in Chapter 7). Shibata (1999) notes that estimation error of these two matrices can cause instability of the results of model selection. Consider the case where a candidate model has $K = 20$ parameters. Then the matrices $J(\theta)$ and $I(\theta)$ are of dimension 20×20, and reliable estimation of the elements of each matrix will be difficult unless sample size is very large. It turns out that $\text{tr}(J(\theta)I(\theta)^{-1})$ itself has a very simple parsimonious estimator, namely K. This is an interesting and important general result.

Thus, AIC is an approximation to TIC, where $\text{tr}(J(\theta)I(\theta)^{-1}) \approx K$. The approximation is excellent when the approximating model is "good" and becomes poor when the approximating model is a poor. However, for models that are poor, the first term, $-2\log(\mathcal{L})$, dominates the criterion because the fit is poor and this term will tend to be relatively large, compared to any much better model. Thus, with the final approximation that $\text{tr}(J(\theta)I(\theta)^{-1}) \approx K$, one can see that AIC is an asymptotically unbiased estimator of relative, expected K-L information, derived without assuming that full reality exists as a model or that such a model is in the set of candidate models. While TIC is an important contribution to the literature, it has rarely seen application. We do not

recommend its use, unless sample size is very large and good estimates of the elements of the matrices $J(\theta)$ and $I(\theta)$ can be expected. Even when this can be done, we expect $\text{tr}(J(\theta)I(\theta)^{-1})$ to be very close to K.

2.4 Second-Order Information Criterion: 1978

While Akaike derived an estimator of K-L information, AIC may perform poorly if there are too many parameters in relation to the size of the sample (Sugiura 1978, Sakamoto et al. 1986). Sugiura (1978) derived a second-order variant of AIC that he called c-AIC.

A Small Sample AIC

Hurvich and Tsai (1989) further studied this small-sample (second-order) bias adjustment, which led to a criterion that is called AIC_c,

$$\mathbf{AIC_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K\left(\frac{n}{n-K-1}\right),}$$

where the penalty term is multiplied by the correction factor $n/(n-K-1)$. This can be rewritten as

$$\mathbf{AIC_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K + \frac{2K(K+1)}{n-K-1},}$$

or, equivalently,

$$\mathbf{AIC_c = AIC + \frac{2K(K+1)}{n-K-1},}$$

where n is sample size (also see Sugiura 1978).

Unless the sample size is large with respect to the number of estimated parameters, use of AIC_c is recommended.

AIC_c merely has an additional bias-correction term. If n is large with respect to K, then the second-order correction is negligible and AIC should perform well. Findley (1985) noted that the study of bias correction is of interest in itself; the exact small-sample bias-correction term varies by type of model (e.g., normal, exponential, Poisson). Bedrick and Tsai (1994) provide a further refinement, but it is more difficult to compute (also see Hurvich and Tsai 1991 and 1995a and b, and Hurvich et al. 1990). While AIC_c was derived under Gaussian assumptions for linear models (fixed effects), Burnham et al. (1994) found this second-order approximation to the K-L distance to be useful in product multinomial models. **Generally, we advocate the use of AIC_c when the ratio n/K is small (say < 40)**. In reaching a decision about the use of AIC vs. AIC_c, one must use the value of K for the highest-dimensioned (i.e., global) model in the set of candidates. If the ratio n/K is sufficiently large, then AIC and AIC_c are similar and will strongly tend to select the same model. One must use either AIC or AIC_c consistently in a given analysis, rather than

mixing the two criteria. Few software packages provide AIC_c values, but these can easily be computed by hand.

2.5 Modification of Information Criterion for Overdispersed Count Data

In general, if the random variable n represents a count under some simple discrete distribution (e.g., Poisson or binomial), it has a known expectation, $\mu(\theta)$, and a known theoretical variance function, $\sigma^2(\theta)$ (θ still is unknown). In a model of overdispersed data the expectation of n is not changed, but the variance model must be generalized, for example using a multiplicative factor, e.g., $\gamma(\theta)\sigma^2(\theta)$. The form of the factor $\gamma(\theta)$ can be partly determined by theoretical considerations and can be complex (see, e.g., McCullagh and Nelder 1989). Overdispersion factors typically are small, ranging from just above 1 to perhaps 3 or 4 if the model structure is correct and overdispersion is due to small violations of assumptions such as independence and parameter homogeneity over individuals. Hence, a first approximation for dealing with overdispersion is to use a simple constant c in place of $\gamma(\theta)$, and this can be generalized to more than one c for different partitions of the data.

Count data have been known not to conform to simple variance assumptions based on binomial or multinomial distributions (e.g., Bartlett 1936, Fisher 1949, Armitage 1957, and Finney 1971). There are a number of statistical models for count data (e.g., Poisson, binomial, negative binomial, multinomial). In these, the sampling variance is theoretically determined, *by assumption* (e.g., for the Poisson model, $\text{var}(n) = \text{E}(n)$; for the binomial model, $\text{var}(\hat{p}) = p(1-p)/n$. If the sampling variance exceeds the theoretical (model-based) variance, the situation is called "overdispersion." Our focus here is on a lack of independence in the data leading to overdispersion, or "extrabinomial variation." Eberhardt (1978) provides a clear review of these issues in the biological sciences. For example, Canada geese (*Branta* species) frequently mate for life, and the pair behaves almost as an individual, rather than as two independent "trials." The young of some species continue to live with the parents for a period of time, which can also cause a lack of independence of individual responses. Further reasons for overdispersion in biological systems include species whose members exist in schools or flocks. Members of such populations can be expected to have positive correlations among individuals within the group; such dependence causes overdispersion. A different type of overdispersion stems from parameter heterogeneity, that is, individuals having unique parameters rather than the same parameter (such as survival probability) applying to all individuals.

The estimators of model parameters often remain unbiased in the presence of overdispersion, but the model-based theoretical variances overestimate precision (McCullagh and Nelder 1989). To properly cope with overdispersion

one needs to model the overdispersion and then use generalized likelihood inference methods. Quasi-likelihood (Wedderburn 1974) theory is a basis for the analysis of overdispersed data (also see Williams 1982, McCullagh and Pregibon 1985, Moore 1987, and McCullagh and Nelder 1989, Lindsey 1999a). Hurvich and Tsai (1995b) provide information on the use of AIC_c with overdispersed data.

Cox and Snell (1989) discuss modeling of count data and note that the first useful approximation is based on a single variance inflation factor (c), which can be estimated from the goodness-of-fit chi-square statistic (χ^2) of the global model and its degrees of freedom,

$$\hat{c} = \chi^2/\mathrm{df}.$$

The variance inflation factor should be estimated from the global model. Cox and Snell (1989) assert that the simple approach of a constant variance inflation factor should often be adequate, as opposed to the much more arduous task of seeking a detailed model for the $\gamma(\theta)$. In a study of these competing approaches on five data sets, Liang and McCullagh (1993) found that modeling overdispersion was clearly better than use of a single $\hat{c}$ in only one of five cases examined.

Given $\hat{c}$, empirical estimates of sampling variances ($\mathrm{var}_e(\hat{\theta}_i)$) and covariances ($\mathrm{cov}_e(\hat{\theta}_i, \hat{\theta}_j)$) can be computed by multiplying the estimates of the theoretical (model-based) variances and covariances by $\hat{c}$ (a technique that has long been used; see, e.g., Finney 1971). These empirical measures of variation (i.e., $\hat{c} \cdot \widehat{\mathrm{var}}_t(\hat{\theta}_i)$) must be treated as having the degrees of freedom used to compute $\hat{c}$ for purposes of setting confidence limits (or testing hypotheses). **The number of parameters (K) must include one for the estimation of c, the variance inflation factor, if used**. Generally, quasi-likelihood adjustments (i.e., use of $\hat{c} > 1$) are made only if some distinct lack of fit has been found (for example, if the observed significance level $P \leq 0.15$ or 0.25) and the goodness-of-fit degrees of freedom ≥ 10, as rough guidelines.

We might expect $c > 1$ with real data but would not expect c to exceed about 4 if model structure is acceptable and only overdispersion is affecting c (see Eberhardt 1978). Substantially larger values of c (say, 6–10) are usually caused partly by a model structure that is inadequate; that is, the fitted model does not account for an acceptable amount of variation in the data. Quasi-likelihood methods of variance inflation are most appropriate only after a reasonable structural adequacy of the model has been achieved. The estimate of c should be computed only for the global model; one should not make and use separate estimates of this variance inflation factor for each of the candidate models in the set. The issue of the structural adequacy of the model is at the very heart of good data analysis (i.e., the reliable identification of the structural versus residual variation in the data). Patterns in the goodness-of-fit statistics (Pearson χ^2 or G-statistics) might be an indication of structural problems with the model. Of course, the biology of the organism in question and the sampling

protocol should provide clues as to the existence of overdispersion; one should not rely only on statistical considerations in this matter.

When data are overdispersed and $c > 1$, the proper likelihood is $\log(\mathcal{L})/c$ (not just $\log(\mathcal{L})$). Principles of quasi-likelihood suggest simple modifications to AIC and AIC_c; we denote these modifications by (Lebreton et al. 1992),

$$\text{QAIC} = -\left[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}\right] + 2K,$$

and

$$\begin{aligned}\text{QAIC}_c &= -\left[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}\right] + 2K + \frac{2K(K+1)}{n-K-1},\\ &= \text{QAIC} + \frac{2K(K+1)}{n-K-1}.\end{aligned}$$

If an overdispersion factor is estimated, then one parameter must be added to K. Of course, when no overdispersion exists, then $c = 1$, and the formulas for QAIC and QAIC_c reduce to AIC and AIC_c, respectively. Anderson et al. (1994) found that these criteria performed well in product multinomial models of capture–recapture data in the presence of differing levels of overdispersion.

One must be careful when using some standard software packages (e.g., SAS GENMOD), since they were developed some time ago under a hypothesis testing mode (i.e., adjusting χ^2 test statistics by $\hat{c}$ to obtain F-tests). In some cases, a separate estimate of c is made for each model, and variances and covariances are multiplied by this model-specific estimate of the variance inflation factor. Some software packages compute an estimate of c for every model, thus making the correct use of model selection criteria tricky unless one is careful. Instead, we recommend that the global model be used as a basis for the estimation of a single variance inflation factor c. Then the empirical

Overdispersed Count Data: A Review

Try to ensure that the structural part of the data is well modeled by the global model.

If there is biological reason to suspect overdispersion, then the overdispersion parameter c can be estimated as χ^2/df, using the global model.

If overdispersion is present, the log-likelihood of the parameter θ, given the data and the model, should be computed as

$$\frac{\log(\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{g}_i))}{\hat{\boldsymbol{c}}}.$$

The number of parameters K is now the number of parameters θ, plus 1 to account for the estimation of the overdispersion parameter c.

The estimated overdispersion parameter should generally be $1 \leq c \leq 4$. Otherwise, some structural lack of fit is probably entering the estimate of overdispersion. If $\hat{c} < 1$, just use $c = 1$.

log-likelihood for each of the candidate models is divided by $\hat{c}$, and QAIC or QAIC_c computed and used for model selection. The estimated variances and covariances should also be adjusted using $\hat{c}$ from the global model, unless there are few degrees of freedom left.

AIC for Overdispersed Count Data

Model selection should use either

$$\mathbf{QAIC} = -[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}] + 2K,$$

or

$$\begin{aligned}\mathbf{QAIC}_c &= -[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}] + 2K + \frac{2K(K+1)}{n-K-1}, \\ &= \mathbf{QAIC} + \frac{2K(K+1)}{n-K-1}\end{aligned}$$

The variance–covariance matrix should be multiplied by the estimated overdispersion parameter $\hat{c}$ (i.e., $\hat{c}(\text{cov}(\hat{\theta}_i, \hat{\theta}_j))$).

Some commercial software computes AIC, while AIC_c is rarely available, and no general software package computes QAIC or QAIC_c. In almost all cases, AIC, AIC_c, QAIC, and QAIC_c can be computed easily by hand from the material that is output from standard computer packages (either likelihood or least squares estimation). In general, we recommend using this extended information-theoretic criterion for count data, and we will use QAIC_c in some of the practical examples in Chapter 3. Of course, often the overdispersion parameter is near 1, negating the need for quasi-likelihood adjustments, and just as often the ratio n/K is large, negating the need for the additional bias-correction term in AIC_c. AIC, AIC_c, and QAIC_c are all estimates of the relative K-L information. We often use the generic term "AIC" to mean any of these criteria.

2.6 AIC Differences, Δ_i

AIC, AIC_c, QAIC_c, and TIC are all on a relative (or interval) scale and are strongly dependent on sample size. Simple differences of AIC values allow estimates of $\text{E}_{\hat{\theta}}[\hat{I}(f, g_i)] - \min \text{E}_{\hat{\theta}}[\hat{I}(f, g_i)]$, where the expectation is over the estimated parameters and min is over the models.

The larger Δ_i is, the less plausible it is that the fitted model $g_i(x|\hat{\theta})$ is the K-L best model, given the data x. Some rough rules of thumb are available and are particularly useful for nested models:

Δ_i	Level of Empirical Support of Model i
0-2	Substantial
4-7	Considerably less
> 10	Essentially none.

AIC Differences

We recommend routinely computing (and presenting in publications) the **AIC differences**,

$$\boldsymbol{\Delta_i = AIC_i - AIC_{min},}$$

over all candidate models in the set. We use the term "AIC differences" in a generic sense here to mean AIC, AIC_c, $QAIC_c$, or TIC. Such differences estimate the relative expected K-L differences between f and $g_i(x|\theta)$. These Δ_i values are easy to interpret and allow a quick comparison and ranking of candidate models and are also useful in computing Akaike weights (Section 2.9). The model estimated to be best has $\Delta_i \equiv \Delta_{min} \equiv 0$.

Models with $\Delta_i > 10$ have either essentially no support, and might be omitted from further consideration, or at least those models fail to explain some substantial explainable variation in the data. These guidelines seem useful if R is small (even as many as 100), but may break down in exploratory cases where there may be thousands of models. The guideline values may be somewhat larger for nonnested models, and more research is needed in this area (e.g., Linhart 1988). If observations are not independent, but are assumed to be independent, then these simple guidelines cannot be expected to hold. Thus, if the log-likelihood is corrected for overdispersion in count data by estimating c, then the guidelines above will be useful.

As an example, candidate models g_1, g_2, g_3, and g_4 have AIC values of 3,400, 3,560, 3,380, and 3,415, respectively. Then one would select model g_3 as the best single model as the basis for inference because g_3 has the smallest AIC value. Because these values are on a relative (interval) scale, one could subtract, say, 3,380 (the minimum of the 4 values) from each AIC value and have the following rescaled AIC values: 20, 180, 0, and 35. Of course, such rescaling does not change the ranks of the models, nor the pairwise differences in the AIC values. People are often surprised that Δ_i of only 1–10 are very important, when the associated AIC values that led to the difference are on the order of 97,000 or 243,000.

AIC Differences

It is not the absolute size of the AIC value, it is the relative values, and particularly the AIC differences (Δ_i), that are important.

An individual AIC value, by itself, is not interpretable due to the unknown constant (interval scale). AIC is only comparative, relative to other AIC values in the model set; thus such differences Δ_i are very important and useful.

We can say with considerable confidence that in real data analysis with several or more models and large sample size (say $n > 10 \times K$ for the biggest model) a model having $\Delta_i = 20$, such as model g_4, would be a very poor approximating model for the data at hand.

We can order the Δ_i from smallest to largest, and the same ordering of the models indicates how good they are as an approximation to the actual, expected K-L best model. Consider Δ_i values for 7 models as 0, 1.2, 1.9, 3.5, 4.1, 5.8, and 7.3. An important question is, how big a difference matters? This should be asked in the sense of when a model is not to be considered competitive with the selected best model as plausibly the actual K-L best model in the set of models used, for the sample size and data at hand. The question has no unambiguous answer; it is like asking how far away from an MLE $\hat{\theta}$ an alternative value of θ must be (assuming that the model is a good model) before we would say that an alternative θ is unlikely as "truth." This question ought to be answered with a confidence (or credibility) interval on θ based on $\hat{\theta}$ and its estimation uncertainty. A conventionally accepted answer here is that θ is unlikely as truth if it is further away than $\pm 2\,\widehat{\text{se}}(\hat{\theta})$ (there is a fundamental basis for using such a procedure). Relative scaling of alternative models can effectively be done using Akaike weights (Section 2.9) and evidence ratios (Section 2.10).

2.7 A Useful Analogy

In some ways, selection of a best approximating model is analogous to auto racing or other similar contests. The goal of such a race is to identify the best (fastest) car/driver combination, and the data represent results from a major race (e.g., the Indianapolis 500 in the USA, the 24 Heures du Mans in France). Only a relatively few car/driver combinations "qualify," based on prerace trials (e.g., 33 cars at Indianapolis)—this is like the set of candidate models (i.e., only certain models "qualify," based on the science of the situation). It would be chaotic if all car/driver combinations with an interest could enter the race, just as it makes little sense to include a very large number of models in the set of candidates (and risk Freedman's paradox). Cars that do not qualify do not win, even though they might indeed have been the best (fastest) had they not failed to qualify. Similarly, models, either good or bad, not in the set of candidates remain out of consideration.

At the end of the race the results provide a ranking ("placing") of each car/driver combination, from first to last. Furthermore, if a quantitative index of quality is available (e.g., elapsed time for each finisher), then a further "scaling" can be considered. Clearly, the primary interest is in "who won the race" or "which was the first"; this is like the model with the minimum AIC value. This answers the question, "Which is best in the race"; the results could differ for another (future) race or another data set, but these are, as yet, unavailable to us.

Some (secondary) interest exists in the question, "Who was in second place?" and in particular, was second place only thousandths of a second behind the winner or 5 minutes behind? The race time results provide answers to these questions, as do the Δ_i values in model selection. In the first case, the best

inference might be that the first two cars are essentially tied and that neither is appreciably better than the other (still, the size of the purse certainly favors the first-place winner!), while in the second case, the inference probably favors a single car/driver combination as the clear best (with a 5-minute lead at the finish). The finishing times provide insights into the third and fourth finishers, etc. In trying to understand the performance of car/driver combinations, one has considerable information from both the rankings and their finishing times, analogous to the AIC values (both the ranks and the Δ_i values). In Sections 2.9 and 2.10 will see how the Δ_i can be used to estimate further quantities, and these will provide additional insights. Note that the absolute time of the winner is of little interest because of temperature differences, track conditions, and other variables; only the *relative* times for a given race are of critical interest. Similarly, the absolute values of AIC are also of little interest, because they reflect sample size and some constants, among other things. The value of the maximized log-likelihood (i.e., $\log(\mathcal{L}(\hat{\theta}|x))$) varies substantially from sample to sample. However, all comparisons of models are made on the same data, so this sample-to-sample variation is irrelevant. Comparing maximized log-likelihood values across data sets is like comparing race finishing times when some races are 500 miles whereas others are 400 or 600 miles.

The winner of the race is clearly the best for the particular race. If one wants to make a broader inference concerning races for an entire year, then results (i.e., ranks) from several races can be pooled or weighted. Similarly, statistical inferences beyond a single observed data set can sometimes be broadened by some type of model averaging using, for example, the nonparametric bootstrap (details in Chapters 4 and 5) and the incorporation of model selection uncertainty in estimators of precision.

The race result might not always select the best car/driver combination, because the fastest qualifying car/driver may have had bad luck (e.g., crash or engine failure) and finished well back from the leader (if at all). Similarly, in model selection one has only one realization of the stochastic process and an *estimated* relative distance as the basis for the selection of a best approximating model (a winner). If the same race is held again with the same drivers, the winner and order of finishers are likely to change somewhat. Similarly, if a new sample of data could be obtained, the model ranks would likely change somewhat.

To carry the analogy a bit further, data dredging would be equivalent to watching a race as cars dropped out and others came to the lead. Then one continually shifts the bet and predicted winner, based on the car/driver in the lead at any point in time (i.e., an unfair advantage). In this case, the final prediction would surely be improved, but the rules of play have certainly been altered! Alternatively, the definition of winning might not be established prior to the initiation of the race. Only after the race are the rules decided (e.g., based, in part, on who they think "ought" to win). Then, one might question the applicability of this specific prediction to other races. Indeed, we recommend "new rules" when data dredging has been done. That is, if a particular result

was found following data dredging, then this should be fully admitted and discussed in resulting publication. We believe in fully examining the data for all the information and insights they might provide. However, the sequence leading to data dredging should be revealed, and results following should be discussed in this light.

Many realize that there is considerable variation in cars and drivers from race to race and track to track. Similarly, many are comfortable with the fact that there is often considerable sampling variation (uncertainty) associated with an estimate of a parameter from data set to data set. Similarly, if other samples (races) could be taken, the estimated best model (car/driver) might also vary from sample to sample (or race to race). Both components of sampling variation and model selection uncertainty should ideally be incorporated into measures of precision.

2.8 Likelihood of a Model, $\mathcal{L}(g_i|data)$

While the AIC differences Δ_i are useful in ranking the models, it is possible to quantify the plausibility of each model as being the actual K-L best model. This can be done by extending the concept of the likelihood of the parameters given both the data and model, i.e., $\mathcal{L}(\theta|x, g_i)$, to the concept of the likelihood of the model given the data, hence $\mathcal{L}(g_i|x)$. Such quantities are very useful in making inferences concerning the relative strength of evidence for each of the models in the set.

Likelihood of a Model, Given Data

The likelihood of model g_i, given the data, is simple to compute for each model in the set:

$$\mathcal{L}(g_i|x) \propto \exp\left(-\frac{1}{2}\Delta_i\right),$$

where "$\propto$" means "is proportional to." Such likelihoods represent the relative strength of evidence for each model.

Akaike (see, e.g., Akaike 1983b) advocates the above $\exp(-\frac{1}{2}\Delta_i)$ for the relative likelihood of the model, given the MLEs of model parameters based on the same data. Such quantities can also be expressed as

$$C\mathcal{L}(\hat{\theta}|x, g_i)e^{-K},$$

where C is an arbitrary constant.

2.9 Akaike Weights, w_i

2.9.1 Basic Formula

Model Probabilities

To better interpret the relative likelihood of a model, given the data and the set of R models, we normalize the $\mathcal{L}(g_i|x)$ to be a set of positive "Akaike weights," w_i, adding to 1:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} \exp(-\frac{1}{2}\Delta_r)}.$$

The w_i depend on the entire set; therefore, if a model is added or dropped during a post hoc analysis, the w_i must be recomputed for all the models in the newly defined set.

This idea of the likelihood of the model given the data, and hence these model weights, has been suggested for many years by Akaike (e.g., Akaike 1978b, 1979, 1980, 1981b and 1983b; also see Bozdogan 1987 and Kishino et al. 1991) and has been researched some by Buckland et al. (1997). These model weights seemed not to have a name, so we call them *Akaike weights*. This name will herein apply also when we use AIC_c, QAIC, QAIC_c, and TIC. **A given w_i is considered as the weight of evidence in favor of model i being the actual K-L best model for the situation at hand *given* that one of the R models must be the K-L best model of that set of R models.** Hence, given that there are only R models and one of them must be best in this set of models, it is convenient to normalize the relative likelihoods to sum to 1.

For the estimated K-L best model (let this be model g_{min}), $\Delta_{min} = 0$; hence, for that model $\exp(-\frac{1}{2}\Delta_{min}) \equiv 1$. The odds for the i^{th} model actually being the K-L best model are thus $\exp(-\frac{1}{2}\Delta_i)$ to 1, or just the "ratio" $\exp(-\frac{1}{2}\Delta_i)$. It is convenient to reexpress such odds as the set of Akaike weights. The bigger a Δ_i is, the smaller the w_i, and the less plausible is model i as being the actual K-L best model for f based on the design and sample size used. The Akaike weights provide an effective way to scale and interpret the Δ_i values. These weights also have other important uses and interpretations that are given in the following chapters.

In general, likelihood provides a good measure of data-based weight of evidence about parameter values, given a model and data (see, e.g., Royall 1997). We think that this concept extends to evidence about the K-L best model, given a set of models. That is, evidence for the best model is well represented by the likelihood of a model.

2.9.2 An Extension

In the absence of any a prior information (in a Bayesian sense) about which of these models might be the K-L best model for the data at hand we are compelled by a certain aspect of information theory itself (see Jaynes 1957, Jessop 1995). Let τ_i be the prior probability that model i is the K-L best model. Lacking any prior information, we set the τ_i all equal, and hence use $\tau_i \equiv 1/R$. In fact, doing so places all R of the models on an equal footing to be selected as the K-L best model.

If there is prior information or belief, this opens the door to unequal prior probabilities. Ignoring any model redundancy (this subject is deferred to Section 4.6), τ_i is our prior state of information or belief that model g_i, fitted to the data, provides the K-L best model for the design and data at hand. This is a deceptively complex issue, as it relates both to ideas of models as best approximations to truth and to expected model fitting tradeoff of bias versus sampling variances.

To us it seems impossible to have any real prior basis for an informative differential assessment of the τ_i (other than on how the models might be structurally interrelated or partially redundant). Using the maximum entropy principle of Jaynes (1957) we should take the τ_i to represent maximal uncertainty about all unknown aspects of the probability distribution represented by the τ_i. Thus we determine the τ_i that maximize the entropy $-\sum \tau_i \log(\tau_i)$ subject to constraints that express whatever information (in the colloquial sense) we have about the distribution. In the "no information" case the only constraint we have is that $\sum \tau_i = 1$ (plus the essential $0 < \tau_i < 1$). The maximum entropy (hence maximum uncertainty) prior is then $\tau_i \equiv 1/R$. [It takes us too far a field to delve into the aspects of information theory underlying the maximum entropy principle. This principle is fundamentally tied both to Boltzmann's entropy and to information theory and can be used to justify noninformative Bayesian priors—when they exist. The interested reader is referred to Kapur and Kesavan 1992, or the less technical Jessop 1995.]

Given any set of prior probabilities (the τ_i), generalized Akaike weights are given by

$$w_i = \frac{\mathcal{L}(g_i|\underline{x})\tau_i}{\sum_{r=1}^{R} \mathcal{L}(g_r|\underline{x})\tau_r}.$$

There may be occasions to use unequal prior probabilities, hence the expression above. However, in general, by Akaike weights we mean the simple expression without the τ_i (this assumes $\tau_i = 1/R$).

The inclusion of prior probabilities (τ_i) in the w_i is not a true Bayesian approach. The full Bayesian approach to model selection requires both the prior τ_i on the model and a prior probability distribution on the parameters θ in model g_i for each model. Then the derivation of posterior results requires integration (usually achievable only by Markov chain Monte Carlo methods). Persons

wishing to learn the Bayesian approach to model selection can start with the following sources: Raftery et al. (1993), Madigan and Raftery 1994, Carlin and Chib (1995), Chatfield (1995b), Draper (1995), Gelman et al. (1995), Kass and Raftery (1995), Hoeting and Ibrahim (1996), Raftery (1996a, 1996b), and Morgan (2000).

A brief comparison is given here of what we mean by the prior probabilities τ_i under this information-theoretic approach to model selection versus what seems to be meant by the prior probabilities of models in the Bayesian approach. The Bayesian approach seems generally to assume that one of the models, in the set of R models, is true. Hence, τ_i is then the prior degree of belief that model form g_i is the true model form (see, e.g., Newman 1997). Under the information-theoretic approach we do not assume that truth f is in the set of models, and $\tau_1, \ldots, \tau_R$ is a probability distribution of our prior information (or lack thereof) about which of the R models is the K-L best model for the data. Information theory itself (Kapur and Kesavan 1992) then justifies determination of the τ_i, generally as $\tau_i \equiv 1/R$. For data analysis we believe that the issue cannot be which model structure is truth, because none of the models considered is truth. Rather, the issue is, which model *when fit to the data* (i.e., when θ is *estimated*) is the best model for purposes of representing the (finite) information in the data. Letting $\tau_i = \text{Prob}\{\text{belief that model form } g_i \text{ is the K-L best model}\}$, then τ_i is about the "parameter" g_{best}, not about the random variable g_{min}. Here, we use only $\tau_i = 1/R$.

2.10 Evidence Ratios

Using the hypothetical example in Section 2.6, the likelihood of each model, given the data, and the Akaike weights are given below:

Model	Δ_i	$\mathcal{L}(g_i\|x)$	Akaike weight w_i
1	0	1	0.431
2	1.2	0.54881	0.237
3	1.9	0.38674	0.167
4	3.5	0.17377	0.075
5	4.1	0.12873	0.056
6	5.8	0.05502	0.024
7	7.3	0.02599	0.010.

As weight of evidence for each model we can see that the selected best model is not convincingly best; the evidence ratio for model g_1 versus model g_2 is only about 2 (i.e., $w_1/w_2 = 1.82$). This relatively weak support for the best model suggests that we should expect to see a lot of variation in the selected best model from sample to sample if we could, in this situation, draw multiple independent samples; that is, the model selection uncertainty is likely to be high. The evidence ratio for the best model versus model 6 is

Evidence Ratios

Evidence can be judged by the relative likelihood of model pairs as

$$\mathcal{L}(g_i|x)/\mathcal{L}(g_j|x)$$

or, equivalently, the ratio of Akaike weights w_i/w_j. Such ratios are commonly used, and we will term them **evidence ratios**. Such ratios represent the evidence about fitted models as to which is better in a K-L information sense.

In particular, there is often interest in the ratio w_1/w_j, where model 1 is the estimated best model and j indexes the rest of the models in the set. These ratios are not affected by any other model, hence do not depend on the full set of R models—just on models i and j. These evidence ratios are invariant to all other models besides i and j.

$0.431/0.024 = e^{(5.8/2)} = 18$, and we must conclude that is it unlikely that model 6 is the K-L best model; the evidence here is reasonably strong against model 6.

There is a striking nonlinearity in the evidence ratios as a function of the Δ_i values. Consider the ratio $w_1/w_j(\equiv w_{min}/w_j)$,

$$\frac{w_1}{w_j} \equiv \frac{1}{e^{-1/2\Delta_j}} \equiv e^{1/2\Delta_j}$$

in the comparison of the evidence for the best model versus the jth best model. Then, we have the following table:

Δ_j	Evidence ratio
2	2.7
4	7.4
8	54.6
10	148.4
15	1,808.0
20	22,026.5

This information helps to justify the rough rules of thumb given for judging the evidence for models being the best K-L model in the set. Jeffreys (1948) provided some likelihood-based rules similar to these over 50 years ago. See Edwards (1992) and Royall (1997) for additional perspectives on the concept of evidence in a likelihood framework.

People may, at first, be frustrated that they do not have some value or cutoff point that provides a simple dichotomy to indicate what is *important* (i.e., "significant" under the Neyman–Pearson null hypothesis testing procedure where a decision is to be reached). Even knowing that statistical significance is not particularly related to biological significance, and that the α-level is arbitrary, some investigators seem to feel comfortable being "told" what is

important. This is the blind hope that the computer and its analysis software will somehow "tell" the investigator what is *important* in a yes or no sense. The approach we advocate is one of quantitative evidence; then people may interpret the quantitative evidence.

Consider a football game where the final score is 10 to 13 for teams A and B, respectively. Here, one does not ask whether the win of team B over team A was "significant." Rather, one can see that the game was close, based on the score (the evidence). Further scrutiny of the evidence could come from examining the total yards gained, the cumulative time of possession of the ball, the number of penalties, etc., for each team. Based on the totality of the evidence, one can reach a determination concerning the relative strength of the two teams. Furthermore, in this case, most rational people will reach roughly the same determination, based on the evidence. Similarly, if the score had been 40 to 3 (the evidence), it would be clear that team A hammered its hapless opponent. Even in this case there is no concept of "highly significant," much less any test of the null hypothesis based on the observed scores that the teams were of equal ability. Again, most rational people would probably agree that team A was the better team on the day of the contest, based on the evidence (40 vs. 3). Based on the evidence, people might be willing to make an inference to other games between these two teams. Of course, there are intermediate cases (10 vs. 16) where the evidence is not convincing. Perhaps the final touchdown occurred in overtime, in which case people might often interpret the evidence (10 to 16) differently. Again, a review of other game statistics might provide insights, but we should admit that not all evidence will lead to a clear determination, accepted by all. One encounters various forms of numerical evidence in everyday life and can interpret such evidence without arbitrary dichotomies.

When we learn that model g_4 has an evidence ratio of 3 in relation to model g_2, it means there is relatively little evidence in favor of model g_4. An analogy here is an auditorium containing N people (let N be large, but unspecified). Each person has a raffle ticket, except that a single person (Bob) has 3 tickets. The evidence ratio (relative likelihood) of Bob winning the raffle vs. any other individual is 3. Clearly, Bob has an edge over any other individual, but it is not strong. Of course, the probability that either Bob or any other particular individual will win is small if N is large. However, the ratio 3/1 remains the same, regardless of the value of N. In contrast, let Bob now have 100 tickets. Then his relative likelihood of winning vs. any other individual is 100, and this is relatively strong evidence. Such evidence ratios are only relative (i.e., Bob vs. another individual); nothing is to be inferred about Bob's chances (or any other individual's chances) of winning the raffle outright. Only Bob's chances relative to another individual's chances are quantified using evidence ratios. Finally, note that the probability of Bob winning, *given that either Bob or another single individual wins*, is $100/(100 + 1) = 0.99$. Evidence ratios for model pairs (e.g., model g_4 vs. model g_2) are relative values.

2.11 Important Analysis Details

Data analysis involves the proper tradeoff between bias and variance or, similarly, between underfitting and overfitting. The estimation of expected K-L information is a natural and simple way to view model selection; given a good set of candidate models, select that fitted model where information loss is minimized. Proper model selection is reflected in good achieved confidence interval coverage for the parameters in the model (or for prediction); otherwise, perhaps too much bias has been accepted in the tradeoff to gain precision, giving a false sense of high precision. This represents the worst inferential situation: a highly precise, but quite biased estimate. These ideas have had a long history in statistical thinking.

An information criterion (i.e., AIC, AIC_c, QAIC, and TIC) can be used to rank the candidate models from best to worst and scale the models using Akaike weights and evidence ratios. Often data do not support only one model as clearly best for data analysis. Instead, suppose three models are essentially tied for best, while another, larger, set of models is clearly not appropriate (either underfit or overfit). Such virtual "ties" for the best approximating model must be carefully considered and admitted. Poskitt and Tremayne (1987) discuss a "portfolio of models" that deserve final consideration. Chatfield (1995b) notes that there may be more than one model that is to be regarded as "useful."

Ambivalence

The inability to ferret out a single best model is not a defect of AIC or any other selection criterion. Rather, it is an indication that the data are simply inadequate to reach such a strong inference. That is, the data are ambivalent concerning some effect or parametrization or structure.

In such cases, all the models in the set can be used to make robust inferences: multimodel inference.

It is perfectly reasonable that several models would serve nearly equally well in approximating the information in a set of data. Inference must admit that there are sometimes competing models and the data do not support selecting only one. The issue of competing models is especially relevant in including model selection uncertainty into estimators of precision. When more than one model has substantial support, some form of multimodel inference (e.g., model averaging) should be considered (Chapter 4). The following subsections provide some important details that must be considered in a careful analysis of research data.

2.11.1 AIC Cannot Be Used to Compare Models of Different Data Sets

Models can be compared using the various information criteria, as estimates of relative, expected K-L information, only when they have been fitted to exactly the same set of data. For example, if nonlinear regression model g_1 is fitted to a

data set with $n = 140$ observations, one cannot validly compare it with model g_2 when 7 outliers have been deleted, leaving only $n = 133$. Furthermore, AIC cannot be used to compare models where the data are ungrouped in one case (Model U) and grouped (e.g., grouped into histograms classes) in another (Model G).

Data Must Be Fixed

An important issue, in general, is that the data and their exact representation must be fixed and alternative models fitted to this fixed data set.

Information criteria should not be compared across different data sets, because the inference is conditional on the data in hand.

2.11.2 Order Not Important in Computing AIC Values

The order in which the information criterion is computed over the set of models is not relevant. Often, one may want to compute AIC_c, starting with the global model and proceed to simpler models with fewer parameters. Others may wish to start with the simple models and work up to the more general models with many parameters; this strategy might be best if numerical problems are encountered in fitting some high-dimensioned models. The order is irrelevant here to proper interpretation, as opposed to the various hypothesis testing approaches where the order may be both arbitrary and the results quite dependent on the choice of order (e.g., stepup (forward) vs. stepdown (backward) testing; Section 3.4.6 provides an example).

2.11.3 Transformations of the Response Variable

Model selection methods assume that some response variable (say y) is the subject of interest. Assuming that the scientific hypotheses relate to this response variable, then all the models must represent exactly this variable. Thus, the R models in the set should all have the same response variable. A common type of mistake is illustrated by the following example. An investigator is interested in modeling a response variable y and has built 4 linear regression models of y, but during the model building, he decides to include a nonlinear model. At that point he includes a model for $\log(y)$ as the fifth model. Estimates of K-L information in such cases cannot be validly compared. This is an important point, and often overlooked. In this example, one would find g_5 to be the best model followed by the other 4 models, each having large Δ_i values. Based on this result, one would erroneously conclude the importance of the nonlinearity. **Investigators should be sure that all hypotheses are modeled using the same response variable (e.g., if the whole set of models were based on log(y), no problem would be created; it is the mixing of response variables that is incorrect).**

Elaborating further, if there was interest in the normal and log-normal model forms, the models would have to be expressed, respectively, as,

$$g_1(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{[y-\mu]^2}{\sigma^2}\right],$$

and another model,

$$g_2(y|\mu,\sigma) = \frac{1}{y\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{[\log(y)-\mu]^2}{\sigma^2}\right].$$

Another critical matter here is that all the components of each likelihood should be retained in comparing different probability distributions. There are some comparisons of different pdfs in this spirit in Section 6.7.1. This "retain it all" requirement is not needed in cases like multiple regression with constant variance because all the comparisons are about the model structure (i.e., variables to select) with an assumption of normal errors for every model. In this case there is a global model and its associated likelihood, and the issue is how best to represent μ as a regression function.

In other cases, it is tempting to drop constants in the log-likelihood, because they do not involve the model parameters. However, alternative models may not have the same constants; this condition makes valid model comparisons impossible. The simple solution here is to retain all the terms in the log-likelihood for all the models in the set.

2.11.4 Regression Models with Differing Error Structures

This issue is related to that in Section 2.11.3. A link between the residual sum of squares (RSS) and σ^2 from regression models with normally distributed errors to the maximized log-likelihood value was provided in Section 1.2.2. This link is a special case, allowing one to work in an ordinary least squares regression framework for modeling and parameter estimation and then switch to a likelihood framework to compute $\log(\mathcal{L}(\theta|data,\ model))$ and various other quantities under an information-theoretic paradigm.

The mapping from $\hat{\sigma}^2$ to $\log(\mathcal{L}(\theta|data,\ model))$ is valid only if all the models in the set assume independent, normally distributed errors (residuals) with a constant variance. If some subset of the R models assume lognormal errors, then valid comparisons across all the models in the set are not possible. In this case, all the models, including those with differing error structures, should be put into a likelihood framework since this permits valid estimates of $\log(\mathcal{L}(\theta|data,\ model))$ and criteria such as AIC_c.

2.11.5 *Do Not Mix Null Hypothesis Testing with Information-Theoretic Criteria*

Tests of null hypotheses and information-theoretic approaches should not be used together; they are very different analysis paradigms. A very common mistake seen in the applied literature is to use AIC to rank the candidate models and then "test" to see whether the best model (the alternative hypothesis) is "significantly better" than the second-best model (the null hypothesis). This procedure is flawed, and we strongly recommend against it (Anderson et al. 2001*c*). Despite warnings about the misuse of hypothesis testing (see Anderson et al. 2000, Cox and Reid 2000), researchers are still reporting P-values for trivial null hypotheses, while failing to report effect size and its precision.

Some authors state that the best model (say g_3) is *significantly* better than another model (say g_6) based on a Δ value of 4–7. Alternatively, sometimes one sees that model g_6 is *rejected* relative to the best model. These statements are poor and misleading. It seems best not to associate the words *significant* or *rejected* with results under an information-theoretic paradigm. Questions concerning the strength of evidence for the models in the set are best addressed using the evidence ratio (Section 2.10), as well as an analysis of residuals, adjusted R^2, and other model diagnostics or descriptive statistics.

2.11.6 *Null Hypothesis Testing Is Still Important in Strict Experiments*

A priori hypothesis testing plays an important role when a formal experiment (i.e., treatment and control groups being formally contrasted in a replicated design with random assignment) has been done and specific a priori alternative hypotheses have been identified. In these cases, there is a very large body of statistical theory on testing of treatment effects in such experimental data. We certainly acknowledge the value of traditional testing approaches to the analysis of these *experimental* data. Still, the primary emphasis should be on the size of the treatment effects and their precision; too often we find a statement regarding "significance," while the treatment and control means are not even presented (Anderson et al. 2000 Cox and Reid 2000). Nearly all statisticians are calling for estimates of effect size and associated precision, rather than test statistics, P-values, and "significance."

Akaike (1981) suggests that the "multiple comparison" of several treatment means should be viewed as a model selection problem, rather than resorting to one of the many testing methods that have been developed (also see Berry 1988). Here, a priori considerations would be brought to bear on the issue and a set of candidate models derived, letting information criterion values aid in sorting out differences in treatment means—a refocusing on parameter estimation, instead of on testing. An alternative approach is to consider random effects modeling (Kreft and deLeeuw 1998).

In observational studies, where randomization or replication is not achievable, we believe that "data analysis" should be viewed largely as a problem in model selection and associated parameter estimation. This seems especially the case where nuisance parameters are encountered in the model, such as the recapture or resighting probabilities in capture–recapture or band–recovery studies. Here, it is not always clear what either the null or the alternative hypothesis should be in a hypothesis testing framework. In addition, often hypotheses that are tested are naive or trivial, as Johnson (1995, 1999) points out with such clarity. Should we expend resources to find out if ravens are white? Is there any reason to test formally hypotheses such as "H_0: the number of robins is the same in cities A and B"? Of course not! One should merely assume that the number is different and proceed to estimate the magnitude of the difference and its precision: an estimation problem, not a null hypothesis testing problem.

2.11.7 Information-Theoretic Criteria Are Not a "Test"

The theories underlying the information-theoretic approaches and null hypothesis testing are fundamentally quite different.

Criteria Are Not a Test

Information-theoretic criteria such as AIC, AIC_c, and $QAIC_c$ are not a "test" in any sense, and there are no associated concepts such as test power or P-values or α-levels. Statistical hypothesis testing represents a very different, and generally inferior, paradigm for the analysis of data in complex settings.

It seems best to avoid use of the word "significant" in reporting research results under an information-theoretic paradigm.

The results of model selection under the two approaches might happen to be similar with simple problems; however, in more complex situations, with many candidate models, the results of the two approaches can be quite different (see Section 3.5). **It is critical to bear in mind that there is a theoretical basis to information-theoretic approaches to model selection criteria, while the use of null hypothesis testing for model selection must be considered ad hoc** (albeit a very refined set of ad hoc procedures in some cases).

2.11.8 Exploratory Data Analysis

Hypothesis testing is commonly used in the early phases of exploratory data analysis to iteratively seek model structure and understanding. Here, one might start with 3–8 models, compute various test statistics for each, and note that several of the better models each have a gender effect. Thus, additional models are generated to include a gender effect, and more null hypothesis tests are conducted. Then the analyst notes that several of these models have a trend in time for some set of estimable parameters; thus more models with this effect are generated, and so on. While this iterative or sequential strategy violates

several theoretical aspects of hypothesis testing, it is very commonly used, and the results are often published without the details of the analysis approach. We suggest that if the results are treated only as alternative hypotheses for a more confirmatory study to be conducted later, this might be an admissible practice, particularly if other information is incorporated during the design stage. Still, the sequential and arbitrary nature of such testing procedures make us wonder whether this is really a good exploratory technique because it too readily keys in on unique features of the sample data at hand (see Tukey 1980). In any event, the key here is to conduct further investigations based partially on the "hunches" from the tentative exploratory work. Conducting the further investigation has too often been ignored and the tentative "hunches" have been published as if they were a priori results. Often, the author does not admit to the post hoc activities that led to the supposed results.

We suggest that information-theoretic approaches might serve better as an exploratory tool; at least key assumptions upon which these criteria are based are not terribly violated, and there is no arbitrary α level. Exploratory data analysis using an information-theoretic criterion, instead of some form of test statistic, eliminates inferential problems in interpreting the many P-values, but one must still worry about overfitting and spurious effects (Anderson et al. 2001b). The ranking of alternative models (the Δ_i and w_i values) might be useful in the preliminary examination of data resulting from a pilot study. Based on these insights, one could design a more confirmatory study to explore the issue of interest. The results of the pilot exploration should remain unpublished. While we do not condone the use of information theoretic approaches in blatant data dredging, we suggest that it might be a more useful tool than hypothesis testing in exploratory data analysis where little a priori knowledge is available. Data dredging has enough problems and risks without using a testing-based approach that carries its own set of substantial problems and limitations.

2.12 Some History and Further Insights

Akaike (1973) considered AIC and its information theoretic foundations ". . . a natural extension of the classical maximum likelihood principle." Interestingly, Fisher (1936) anticipated such an advance over 60 years ago when he wrote,

> . . . an even wider type of inductive argument may some day be developed, which shall discuss methods of assigning from the data the functional form of the population.

This comment was quite insightful; of course, we might expect this from R. A. Fisher! Akaike was perhaps kind to consider AIC an extension of classical ML theory; he might just as well have said that classical likelihood theory was a special application of the more general information theory. In fact, Kullback believed in the importance of information theory as a unifying principle in statistics.

2.12.1 Entropy

Akaike's (1977) term "*entropy maximization principle*" comes from the fact that the negative of K- L information is Boltzmann's entropy (in fact, K-L information has been called negative entropy or "negentropy"). Entropy is "disorder," while max entropy is maximum disorder or minimum information. Conceptually,

$$\text{Boltzmann's entropy } = -\log\left(\frac{f(x)}{g(x)}\right).$$

Then,

$$-\text{Boltzmann's entropy } = \log\left(\frac{f(x)}{g(x)}\right),$$

and

$$\begin{aligned}\text{K-L } &= \mathrm{E}_f(-\text{Boltzmann's entropy})\\ &= \mathrm{E}_f\left(\log\left(\frac{f(x)}{g(x)}\right)\right),\\ &= \int f(x)\log\left(\frac{f(x)}{g(x)}\right)\,dx.\end{aligned}$$

Thus, minimizing the K-L distance is equivalent to maximizing the entropy; hence the name *maximum entropy principle* (see Jaynes 1957, Akaike 1983a, 1985 and Bozdogan 1987, Jessop 1995 for further historical insights). However, maximizing entropy is subject to a constraint—the model of the information in the data. A good model contains the information in the data, leaving only "noise." It is the noise (entropy or uncertainty) that is maximized under the concept of the entropy maximization principle (Section 1.2.4). Minimizing K-L information then results in an approximating model that loses a minimum amount of information in the data. Entropy maximization results in a model that maximizes the uncertainty, leaving only information (the model) "maximally" justified by the data. The concepts are equivalent, but minimizing K-L distance (or information loss) certainly seems the more direct approach.

The K-L information is *averaged* negative entropy, hence the expectation with respect to f. While the theory of entropy is a large subject by itself, readers here can think of entropy as nearly synonymous with uncertainty, or randomness or disorder in physical systems.

Boltzmann derived the fundamental theorem that

entropy is proportional to − log(probability).

Entropy, information, and probability are thus linked, allowing probabilities to be multiplicative while information and entropies are additive. (This result was also derived by Shannon 1948). Fritz Hasenöhrl, a student of Boltzmann, Boltzmann's successor at Vienna University, and a famous theoretical

physicist himself, noted that this result ". . . is one of the most profound, most beautiful theorems of theoretical physics, indeed all of science." Further information concerning Boltzmann appears in Brush (1965, 1966), while interesting insights into Akaike's career are found in Findley and Parzen (1995).

2.12.2 A Heuristic Interpretation

After Akaike's innovative derivation of AIC, people noticed a heuristic interpretation that was both interesting and sometimes misleading. The first term in AIC,

$$\text{AIC} = -2\log(\mathcal{L}(\hat{\theta}|x)) + 2K,$$

is a measure of lack of model fit, while the second term ($2K$) can be interpreted as a "penalty" for increasing the size of the model (the penalty enforces parsimony in the number of parameters). This heuristic explanation does not do justice to the much deeper theoretical basis for AIC (i.e., the link with K-L distance and information theory). The heuristic interpretation led some statisticians to consider "alternative" penalty terms, and this has not always been productive (see Chapter 6). The so-called penalty term in AIC is not arbitrary; rather, it is the asymptotic bias-correction term. It is the result of deriving an asymptotic estimator of relative, expected K-L information. [Note, of course, that had Akaike defined AIC $= -\log(\mathcal{L}(\hat{\theta}|x)) + K$, the minimization would be unchanged; some authors use this expression, but we will use AIC as Akaike defined it.]

The heuristic view of the components of AIC clearly shows a bias vs. variance tradeoff and insight into how the principle of parsimony is met by using AIC (see Gooijer et al. 1985:316). Still, we recommend viewing AIC as an estimate of the relative expected K-L information or distance between model pairs (i.e., each g_i vs. f). Minimizing this relative, expected distance provides an estimated best approximating model for that particular data set (i.e., the *closest* approximating model to f). The relative K-L distance is the link between information theory and the log-likelihood function that is a critical element in AIC model selection.

2.12.3 More on Interpreting Information-Theoretic Criteria

Estimates of relative K-L information, the AIC differences (Δ_i), or the Akaike weights (w_i) provide a ranking of the models; thus the analyst can determine which fitted model is best, which are essentially tied for best, and which models are clearly in an inferior class (and perhaps some that are in an intermediate class). These ranks are, of course, estimates based on the data. Still, the rankings are quite useful (cf. Section 2.7 and Sakamoto et al. 1986:84) and suggest that primary inference be developed using the model for which AIC is minimized or the small number of models where there is an essential tie for the minimum

AIC (i.e., within about 1 or 2 AIC units from the minimum for nested models successively differing by one parameter). In the context of a string of nested models, when there is a single model that is clearly superior (say, the next best model is $>$ 9–10 AIC units from the minimum) there is little model selection uncertainty and the theoretical standard errors can be used (e.g., Flather's data in Sections 1.2.3 and 2.14). When the results of model selection are less clear, then methods described in Chapter 4 can be considered. AIC allows a ranking of models and the identification of models that are nearly equally useful versus those that are clearly poor explanations for the data at hand (e.g., Table 2.2). Hypothesis testing provides no general way to rank models, even for models that are nested.

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 set 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 may well 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 allow insights into model uncertainty; this and other similar issues are the subject of some of the following chapters.

2.12.4 Nonnested Models

A substantial advantage in using information-theoretic criteria is that they are valid for nonnested models (e.g., Table 2.2). Of course, traditional likelihood ratio tests are defined only for nested models, and this represents another substantial limitation in the use of hypothesis testing in model selection. The ranking of models using AIC helps clarify the importance of modeling (Akaike 1973:173); for example, some models for a particular data set are simply poor and should not be used for inference.

A well-thought-out global model (where applicable) is very important, and substantial prior knowledge is required during the entire survey or experiment, including the clear statement of the question to be addressed and the collection of the data. This prior knowledge is then carefully input into the development of the set of candidate models (Section 1.2.4). Without this background science, the entire investigation should probably be considered only very preliminary.

2.12.5 Further Insights

Much of the research on model selection has been in regression and time series models, with some work being done in log-linear and classical multivariate (e.g., factor analysis) models. Bozdogan (1987) provides a review of the theory and some extensions. However, the number of published papers that critically examine the performance of AIC-selected models is quite limited. One serious problem with the statistical literature as regards the evaluation of AIC has been the use of Monte Carlo methods using only very simple generating models with a few large effects and no smaller, tapering effects. Furthermore, these Monte Carlo studies usually have a poor objective, namely, to evaluate how often a criterion selects the simple generating model. We believe that this misses the point entirely with respect to real data analysis. Such evaluations are often done even without regard for sample size (and often use AIC when AIC_c should have been used).

In Monte Carlo studies it would be useful to generate data from a much more realistic model with several big effects and a series of smaller, tapering effects (Speed and Yu 1993). Then interest is refocused onto the selection of a good approximating model and its statistical properties, rather than trying to select the simple, artificial model used to generate the data. AIC attempts to select a best approximating model for the data at hand; if (as with reality) the "true model" is at all complex, its use, with estimated parameters rather than true ones, would be poor for inference, even if it existed and its functional form (but not parameter values) were known (e.g., Sakamoto et al. 1986). This counterintuitive result occurs because the (limited) data would have to be used to estimate all the unknown parameters in the "true model," which would likely result in a substantial loss of precision (see Figure 1.3B).

AIC reformulates the problem explicitly as a problem of *approximation* of the true structure (probably infinite-dimensional, at least in the biological sciences) by a *model*. Model selection then becomes a simple function minimization, where AIC (or more properly K-L information loss) is the criterion to be minimized. AIC selection is objective and represents a very different paradigm to that of null hypothesis testing and is free from the arbitrary α levels, the multiple-testing problem, and the fact that some candidate models might not be nested. The problem of what model to use is inherently not a hypothesis testing problem (Akaike 1974). However, the fact that AIC allows a simple comparison of models does not justify the comparison of all possible models (Akaike 1985 and Section 1.3.3). If one had 10 variables, then there would be 1,024 possible models, even if interactions and squared or cubed terms are excluded. If sample size is $n \leq 1{,}000$, overfitting the data is almost a certainty. It is simply not sensible to consider such a large number of models, because a model that overfits the data will almost surely result, and the science of the problem has been lost. *Even in a very exploratory analysis it seems poor practice to consider all possible models; surely, some science can be brought to bear on such an unthinking approach* (otherwise, the scientist is superfluous and the work could be done by a technician).

2.13 Bootstrap Methods and Model Selection Frequencies π_i

The bootstrap is a type of Monte Carlo method used frequently is applied statistics. This computer-intensive approach is based on resampling of the observed data (Efron and Tibshirani 1993, Mooney and Duval 1993). The bootstrap was first described by Bradley Efron (1979); thousands of papers have been written on the bootstrap, with various extensions and applications in the past two decades, and it has found very wide use in applied problems. The bootstrap can be used for several purposes, particularly in the robust estimation of sampling variances or standard errors and (asymmetrical) confidence intervals. It has been used in the estimation of model selection frequencies (π_i) and in estimates of precision that include model selection uncertainty.

The bootstrap has enormous potential for the biologist with programming skills; however, its computer intensive nature will continue to hinder its use for large problems. We believe that at least 1,000 bootstrap samples are needed in many applications, and often 10,000 samples are needed for some aspects of model selection. In extreme cases, reliable results could take days of computer time to apply the bootstrap to complex data analysis cases involving large sample size and several dozen models, where the MLEs in each model must be found numerically.

The fundamental idea of the model-based sampling theory approach to statistical inference is that the data arise as a sample from some conceptual probability distribution f. Uncertainties of our inferences can be measured if we can estimate f. The bootstrap method allows the computation of measures of our inference uncertainty by having a simple empirical estimate of f and sampling from this estimated distribution. In practical application, the empirical bootstrap means using some form of resampling *with replacement* from the actual data x to generate B (e.g., $B = 1{,}000$ or 10,000) bootstrap samples; a bootstrap sample is denoted as x_b, where ($b = 1, 2, \ldots, B$). The sample data consist of n independent units, and it then suffices to take a simple random sample of size n, *with replacement*, from the n units of data, to get one bootstrap sample. However, the nature of the correct bootstrap data resampling can be more complex for more complex data structures.

The set of B bootstrap samples is a proxy for a set of B independent real samples from f (in reality we have only one actual sample of data). Properties expected from replicate real samples are inferred from the bootstrap samples by analyzing each bootstrap sample exactly as we first analyzed the real data sample. From the set of results of sample size B we measure our inference uncertainties from sample to (conceptual) population (Figure 2.6). For many applications it has been theoretically shown (e.g., Efron and Gong 1983, Efron and Tibshirani 1993) that the bootstrap can work well for large sample sizes (n), but it is not generally reliable for small n (say 5, 10, or perhaps even 20), regardless of how many bootstrap samples B are used. The bootstrap is not always successful in model selection (see Freedman et al. 1988).

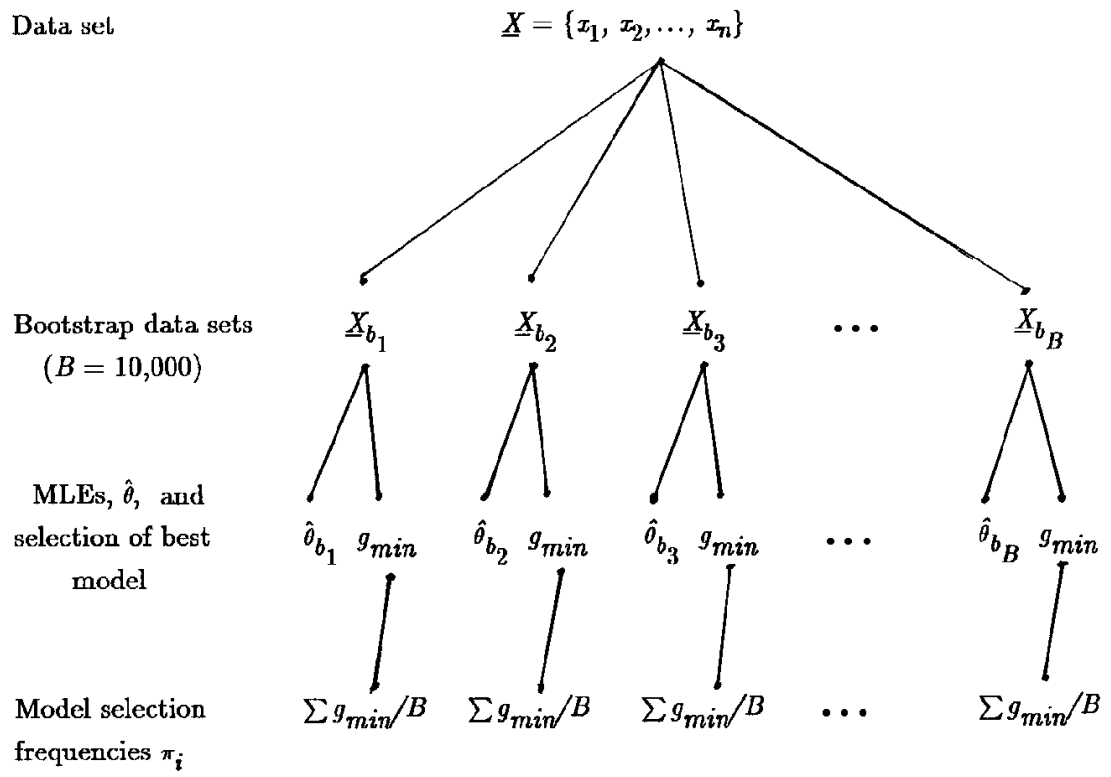

FIGURE 2.6. Diagram of the nonparametric bootstrap method as used in model selection (redrawn from Efron and Tibshirani 1993). The actual data set $\underline{X}$ is sampled with replacement, using the same sample size (n); this is done B times, to obtain B bootstrap data sets $\underline{X}_b$. Maximum likelihood theory provides estimates of the parameters ($\hat{\theta}$) for each of the models i ($i = 1, 2, \ldots, R$) and the AIC-best model (denoted by model $g_{\min}$) is found and its index stored for each of the bootstrap data sets. Finally, the model selection relative frequencies (π_i) are computed as the sums of the frequencies where model i was selected as best, divided by B. Of course, $\sum \pi_i = 1$.

2.13.1 Introduction

In many cases one can derive the sampling variance of an estimator from general likelihood theory. In other cases, an estimator may be difficult to derive or may not exist in closed form. For example, the finite rate of population change (λ) can be derived from a Leslie population projection matrix (a function of age-specific fecundity and age-specific, conditional survival probabilities). Generally, λ cannot be expressed in closed form. The bootstrap is handy for variance estimation in such nonstandard cases.

Consider a sample of weights of 27 young rats ($n = 27$); the data are (from Manly 1992),

57 60 52 49 56 46 51 63 49 57 59 54 56 59 57 52 52 61 59 53 59 51 51 56 58 46 53.

The sample mean of these data is 54.7, and the standard deviation is 4.51 with cv = 0.0824. For illustration, we will estimate of the standard error of the cv. Clearly, this would be nonstandard; however, it represents a way to illustrate the bootstrap.

First, we draw a random subsample of size 27 *with replacement* from the actual data. Thus, while a weight of 63 appears only once in the actual sample, perhaps it would not appear in the subsample; or it could appear more than once. Similarly, there are 3 occurrences of the weight 57 in the actual sample; perhaps the bootstrap sample would have, by chance, no values of 57. The point here is that a random sample of size 27 is taken *with replacement* from the original 27 data values. This is the first bootstrap resample ($b = 1$). From this bootstrap sample, one computes $\hat{\mu} = \bar{x}$, the $\widehat{\text{se}}(\hat{\mu}) = s/\sqrt{27}$, and the $\text{cv} = \widehat{\text{se}}(\hat{\mu})/\hat{\mu}$, and stores that value of cv in memory.

Second, the whole process is repeated B times (where we will let $B =$ 10,000 samples for this example). Thus, we generate 10,000 resample data sets ($\text{b} = 1, 2, 3, \ldots, 10{,}000$) and from *each* of these we compute $\hat{\mu}$, $\widehat{\text{se}}(\hat{\mu})$, and the cv and store the value of the cv.

Third, we obtain the estimated standard error of the cv pertaining to the original sample by taking the standard deviation of the 10,000 cv values (corresponding to the 10,000 bootstrap samples). The process is simple; in this case, the standard error of the cv is 0.00922, or less than 1%.

Confidence intervals can be computed in the usual way, $\text{cv} \pm 2\,\widehat{\text{se}}(\text{cv})$. This gives a 95% interval of (0.0640, 0.1009) for the rat data. However, the sampling distribution may be nonnormal and a more robust interval might be required. Again, the bootstrap provides a simple approach. In this case, one sorts the $B = 10{,}000$ estimates of the cv in ascending order and selects the values that cut off the lower and upper 2.5 percentiles. Thus, the resulting interval might be asymmetric.

In the rat cv, the percentile bootstrap 95% confidence interval is (0.0626, 0.0984). This interval is about the same width as in the traditional approach, but shifted a bit toward 0. Incidentally, the mean of the 10,000 bootstrap samples was 0.0806 (compared to the actual sample cv of 0.0824). Even $B =$ 1,000 is usually adequate for the estimation of the sampling variance or standard deviation; however, good estimates of percentile confidence intervals may require $B = 10{,}000$ in complicated applications.

Just as the analysis of a single data set can have many objectives, the bootstrap can be used to provide insight into a host of questions. For example, for each bootstrap sample one could compute and store the conditional variance–covariance matrix, goodness-of-fit values, the estimated variance inflation factor, the model selected, confidence interval width, and other quantities. Inference can be made concerning these quantities, based on summaries over the B bootstrap samples.

The illustration of the bootstrap on the rat data is called a nonparametric bootstrap, since no parametric distribution is assumed for the underlying process that generated the data. We assume only that the data in the original sample were "representative" and that sample size was not small. The parametric bootstrap is frequently used and allows assessment of bias and other issues. The use of the parametric bootstrap will be illustrated by the estimation of the variance inflation factor $\hat{c}$.

Consider an open population capture–recapture study in a setting where the investigators suspect a lack of independence because of the way that family groups were captured and tagged in the field. Data analysis reveals $\chi^2_{gof}/\text{df} = 3.2$. The investigators suspected some extrabinomial variation, but are surprised by the large estimate of the variance inflation factor $\hat{c}$. They suspect that the estimate is high and decide to use a parametric bootstrap to investigate their suspicion. They realize that the program RELEASE (Burnham et al. 1987) can be used to do Monte Carlo simulations and output a file with the goodness-of-fit statistics.

They input the MLEs from the real data into RELEASE as if they were parameters (ϕ_j and p_j) and use the numbers of new releases in the field data as input. Then the amount of extrabinomial variation (i.e., overdispersion, but called EBV in RELEASE) is specified. In this illustration, let EBV $\equiv$ 1, meaning no overdispersion. They then run 1,000 Monte Carlo samples and obtain the information on the estimated variance inflation factor for each rep. The average of these 1,000 values gives $\hat{\text{E}}(\hat{c})$, and this can be compared to 1, the value used to generate the data. This result provides insight to the investigators on what to do about possible overdispersion in their data. More generally, the investigators could conduct several such studies for a range of EBV and see whether $\text{E}(\hat{c}|\text{EBV}) = \text{EBV}$ and assess any systematic bias in $\hat{c}$ as an estimator of EBV.

This bootstrap is parametric in that parameters were specified (in this case, from the MLEs from real data that were available) and used in a generating model to produce Monte Carlo data. The nonparametric bootstrap does not require parameters nor a model and relies on resampling the original data.

The bootstrap has been used in population biology to set confidence intervals on the median and mean life span. It is conceptually simple and has found very widespread use in applied statistics. Biologists planning a career in research or teaching should be familiar with the bootstrap. There is a very large literature on the bootstrap; see Efron and Tibshirani (1993) for an introduction to the subject and a large list of references. Some valid applications of the bootstrap are tricky (even multiple linear regression), so some care is required in more complex settings!

2.13.2 The Bootstrap in Model Selection: The Basic Idea

Consider the case where data (x) with sample size n are available and $R = 6$ models are under consideration, each representing some scientific hypothesis of interest. Let $B = 10{,}000$ bootstrap data sets, each of size n, and derived by resampling the data with replacement. MLEs of the parameters for each model could be computed for each bootstrap sample. Then AIC_c could be computed for each of the 6 ($i = 1, 2, \ldots, 6$) models and the number of the best model (denote this by r^*, where r^* is the number of the best of the 6 models) and its associated AIC_c value stored for each of the 10,000 bootstrap samples. After 10,000 such analyses, one has the bootstrap frequency of selection for each

of the 6 models. These are called model selection relative frequencies π_i, the relative frequency that model i was found to be best. The relative frequency is given by $\pi_i =$ frequency$/10{,}000$ in this example. Of course, AIC or QAIC$_c$, or TIC could have been used to estimate the π_i.

Relative frequencies for model i being selected as the best model are similar to the Akalke weights, but are not identical. There is no reason, nor need, for the data-based weights of evidence (as the set of w_i) to be the same as the sampling relative frequencies at which the models are selected by an information criteria as being best. In general, likelihood provides a better measure of data-based weight of evidence about parameter values, given a model and data (see, e.g., Royall 1997), and we think that this concept (i.e., evidence for the best model is best represented by the likelihood of a model) rightly extends to evidence about a best model given an a priori set of models.

In our work we have not seen any particular advantage in the bootstrap selection frequencies over the Akaike weights. Considering the programming and computer times required for the computation of the model selection frequencies, we prefer the Akaike weights in general. We present some comparisons in Chapters 4 and 5.

We further elaborate on the interpretation of the Akaike weights as being conceptually different from the sampling-theory-based relative frequencies of model selection. It has has been noted in the literature (e.g., Akaike 1981a, 1994, Bozdogan 1987) that there is a Bayesian basis for interpreting the Akaike weight w_i as being the probability that model g_i is the expected K-L best model given the data (for convenience we usually drop this "expected" distinction and just think of the K-L best model). Once we have accepted the likelihood of model g_i given the data $\mathcal{L}(g_i|x)$, then we can compute the approximate posterior probability that model g_i is the K-L best model if we are willing to specify prior probabilities on the models (note that some Bayesians would consider this approach ad hoc since it is not the full Bayesian approach). That is, we first must specify an a priori probability distribution $\tau_1, \ldots, \tau_R$, which provides our belief that fitted model g_i will be the K-L best model for the data, given the model set. These probabilities τ_i must be specified independent of (basically, prior to) fitting any models to the data.

2.14 Return to Flather's Models

We now extend the example in Chapter 1 where 9 models for the species-accumulation curve for data from Indiana and Ohio were analyzed by Flather (1992, 1996). The simple computation of AIC was done by hand from the regression output from program NLIN in SAS (SAS Institute, Inc. 1985). In this case, apart from a constant that is the same over all models,

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2) + 2K,$$

TABLE 2.1. Summary of nine a priori models of avian species-accumulation curves from the Breeding Bird Survey (from Flather 1992 and 1996). Models are shown, including the number of parameters (K), AIC values, $\Delta_i = \text{AIC}_i - \text{AIC}_{min}$ values, Akaike weights, and adjusted R^2 values for the Indian–Ohio Major Land Resource Area. AIC is computed for each model; the order is not relevant. Here the models are shown in order according to the number of parameters (K). However, this is only a convenience. This elaborates on the example in Table 1.1.

Model	Number of parameters[a]	AIC value	Δ_i	w_i	Adjusted R^2
ax^b	3	227.64	813.12	0.0000	0.962
$a + b\log(x)$	3	91.56	677.04	0.0000	0.986
$a\Big(x/(b+x)\Big)$	3	350.40	935.88	0.0000	0.903
$a(1-e^{-bx})$	3	529.17	1114.65	0.0000	0.624
$a - bc^x$	4	223.53	809.01	0.0000	0.960
$(a+bx)/(1+cx)$	4	57.53	643.01	0.0000	0.989
$a(1-e^{-bx})^c$	4	−42.85	542.63	0.0000	0.995
$a\Big(1-[1+(x/c)^d]^{-b}\Big)$	5	−422.08	163.40	0.0000	0.999
$a[1-e^{-(b(x-c))^d}]$	5	−585.48	0	1.0000	0.999

[a] K is the number of parameters in the regression model plus 1 for σ^2.

where $\hat{\sigma}^2 = \text{RSS}/n$ and K is the number of regression parameters plus 1 (for σ^2). AIC values for the 9 models are given in Table 2.1. The last model is clearly the best approximating model for these data. Values of $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}} = \text{AIC}_i + 585.48$ are also given and allow the results to be more easily interpreted. Here, the second- and third-best models are quickly identified (corresponding to Δ_i values of 163.40 and 542.63, respectively); however, these Δ values are very large, and the inference here is that the final model is clearly the best of the candidate models considered for these specific data. This conclusion seems to be born out by Flather (1992), since he also selected this model based on a careful analysis of residuals for each of the 9 models and Mallows' C_p. The remaining question is whether a still better model might have been postulated with 6 or 7 parameters and increased structure. **Information criteria attempt only to select the best model from the candidate models available; if a better model exists, but is not offered as a candidate, then the information-theoretic approach cannot be expected to identify this new model.**

Adjusted R^2 values are shown in Table 2.2, and while these are useful as a measure of the proportion of the variation "explained," they are not useful in model selection (McQuarrie and Tsai 1998). In the case of Flather's data, the best 4 models all have an adjusted $R^2 \approx 0.99$, prompting one to conclude (erroneously) that all 4 models are an excellent fit to the data. Examination of the Δ_i values shows that models 6, 7 and 8 are incredibly poor, relative to model 9. The evidence ratio for the best model versus the second-best model

is

$$w_9/w_8 = \frac{1}{\exp(-163.4/2)} \approx 3.0 \times 10^{35}.$$

There are additional reasons why adjusted R^2 is poor in model selection; its usefulness should be restricted to description.

2.15 Summary

Ideally, the investigator has a set of "multiple working hypotheses" and has thought hard about the background science of the issue at hand. Then, the science of the matter, experience, and expertise are used to define an a priori set of candidate models, representing each of these hypotheses. **These are important philosophical issues that must receive increased attention**. The research problem should be carefully stated, followed by careful planning concerning the sampling or experimental design. Sample size and other planning issues should be considered fully before the data-gathering program begins.

The basis for the information-theoretic approach to model selection and inference is **Kullback–Leibler information**,

$$\boldsymbol{I(f,g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx.}$$

I*(*f*, *g*) is the "information" lost when the model *g* is used to approximate full reality or truth *f. An equivalent interpretation of $I(f, g)$ is a "distance" from the approximating model g to full truth or reality f. Under either interpretation, we seek to find a candidate model that minimizes $I(f, g)$, over the candidate models. This is a conceptually simple, yet powerful, approach. However, $I(f, g)$ cannot be used directly, because it requires knowledge of full truth or reality and the parameters in the approximating models g_i.

Akaike (1973), in a landmark paper, provided a way to *estimate* relative, expected $I(f, g)$, based on the empirical log-likelihood function. He found that the maximized log-likelihood value was a biased estimate of relative, expected Kullback–Leibler information and that under certain conditions this bias was approximately equal to K, the number of estimable parameters in the approximating model g. His method, *Akaike's information criterion* (AIC), allowed model selection to be firmly based on a fundamental theory and opened to door to further theoretical work. He considered AIC to be an extension of likelihood theory, the very backbone of statistical theory. Shortly thereafter, Takeuchi (1976) derived an asymptotically unbiased estimator of relative, expected Kullback–Leibler information that applies in general (i.e., without the special conditions underlying Akaike's derivation of AIC). His method (TIC for Takeuchi's information criterion) requires large sample sizes to estimate elements of two $K \times K$ matrices in the bias-adjustment term. TIC represents an important conceptual advance and further justifies AIC. Second order (i.e.,

small sample) approximations (AIC_c) were soon offered by Sugiura (1978) and Hurvich and Tsai (1989 and several subsequent papers). The three main approaches to adjusting for this bias (the bias-adjustment term is subtracted from the maximized log-likelihood) are summarized below:

Criterion	Bias adjustment term
AIC	K
AIC_c	$K + \frac{K(K+1)}{n-K-1}$
TIC	$\text{tr}(J(\theta)I(\theta)^{-1} \approx K.$

These information criteria are estimates of relative, expected K-L information and are an extension of Fisher's likelihood theory. AIC and AIC_c are easy to compute, quite effective in many applications, and we recommend their use. When count data are found to be overdispersed, appropriate model selection criteria have been derived, based on quasi-likelihood theory (QAIC and $QAIC_c$). If overdispersion is found in the analysis of count data, the nominal log-likelihood function must be divided by an estimate of the overdispersion ($\hat{c}$) to obtain the correct log-likelihood. Thus, investigators working in applied data analysis have several powerful methods for selecting a "best" model for making inferences from empirical data to the population or process of interest. In practice, one need not assume that the "true model" is in the set of candidates (although this is sometimes mistakenly stated in the technical literature).

The AIC differences (Δ_i) and Akaike weights (w_i) are important in ranking and scaling the hypotheses, represented by models. The evidence ratios (e.g., w_i/w_j) help sharpen the evidence for or against the various alternative hypotheses. All of these values are easy to compute and simple to understand and interpret.

The principle of parsimony provides a philosophical basis for model selection, K-L information provides an objective target based on deep theory, and AIC, AIC_c, $QAIC_c$, and TIC provide estimators of relative, expected K-L information. Objective model selection is rigorously based on these principles. These methods are applicable across a very wide range of scientific hypotheses and statistical models. We recommend presentation of $\log(\mathcal{L}(\hat{\theta}))$, K, the appropriate information criterion (AIC, AIC_c, $QAIC_c$ or TIC), Δ_i, and w_i for various models in research papers to provide full information concerning the evidence for each of the models.