Appendices

Appendix A: Likelihood Theory

Likelihood methods are much more general, far less taught in applied statistics courses, and easier to understand as a concept or procedure than least squares. The material in much of this book relies on an understanding of likelihood theory, and so a very brief introduction is given here. While likelihood methods underlie both frequentist and Bayesian statistics, there are no more than a handful of applied books on this important subject (examples include McCullagh and Nelder 1989; Edwards 1992; Azzalini 1996; Morgan 2000; Severini 2000; Pawitan 2001) and none of these constitute easy reading.

A.1 Likelihood Functions

The first key point is that the likelihood function is a product of probabilities. The concept can be illustrated by considering events (or outcomes) that can be observed (e.g., the number of “heads” observed from flipping a coin n times). The set of these observations constitute the data. Specifically, the data from a coin flipping study are the number of heads (y) and the number of tails (n–y) from n coin flips. The *probability* of such events can be “assigned.” Underlying each time a head is observed is the probability of a head; call this p. Underlying each observation of a “tail” is it’s probability; call this 1–p.

Tacit assumptions have been made; these are often termed “independent and identically distributed, *iid*. It is easy to believe that the outcomes of coin flips are independent. Whether a single coin is flipped n times or n coins are flipped once, surely one outcome does not influence the next. The term *identically distributed* relates to each coin having the same properties; if one coin has the probability of a head as some value p, the others have that same value (this condition is also known as parameter homogeneity). These are important assumptions; for example, unless independence is assumed probabilities are not simply multiplied and log-likelihoods cannot be summed. Both of these *iid* assumptions can fail in many applications in the life sciences (see Sect. 6.2).

Note, the sum of the number of heads and tails = $n = y + (n-y)$. Likewise, the sum of the probabilities is $1 = p + (1-p)$. All the events and their probabilities must be accounted for under basic rules of probability. Assume a coin is flipped 11 times (n = 11) and 7 heads (y) are observed. Then the likelihood function ($\mathcal{L}$) for this could be written as the product (note the order is not important),

$$\mathcal{L} \propto ppppppp(1-p)(1-p)(1-p).$$

Some simple notation allows this to be written in a more useful form (where $\propto$ means "proportional to")

$$\mathcal{L} \propto p^{y}(1-p)^{n-y}$$

or for example

$$\mathcal{L} \propto p^{7}(1-p)^{4}.$$

Now it should be clear that this is the binomial model. The above shows the likelihood function is proportional; its exact form must include the binomial coefficient

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

where

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

and is read "n choose y" and is the number of ways a sample of size y can be selected from a population of size n. The binomial coefficient $\binom{n}{y}$ does not contain the unknown parameter p and is often omitted for estimation of model parameters. The key here is to focus on the fact that likelihood functions are the product of probabilities. These probabilities come from assigning underlying probabilities to observed events (the data). This is formalized as

$$\mathcal{L}(p \mid n \text{ and } y, \text{binomial}) = \binom{n}{y} p^{y}(1-p)^{n-y},$$

and is read "the likelihood of the unknown parameter p, given the data (n and y) and the model (binomial). The likelihood function allows the estimation of unknown parameters, given the data and the model (g). The scientist has data and can assign probabilities underlying the data if the model is given (or can be selected). This paves the way for a way to estimate the value of parameters in the model.

There are important distinctions between the terms probability and likelihood. Likelihood is relative or comparative; likelihood values do not sum or integrate to 1. Likelihoods are not probabilities. Likelihood values are like raffle tickets. If you have 14 tickets and Barney has only one ticket then the likelihood of you winning the raffle, relative to Barney winning, is 14:1. Likelihoods are functions of the unknown parameters (θ), given the data (x) and the model (g); $\mathcal{L}$ ($\theta|x$, g). In contrast, probabilities sum or integrate to one and are absolute. Probability functions and distributions are functions of the data, given the value of the parameters and a model; $p(x \mid \theta, g)$. Both probabilities and likelihoods are conditional on various things. Both quantities are useful in inductive inference, but they are different (even though lay people might use these interchangeably).

Clearly, the likelihood is a function of (only) the unknown parameter (p in this example), given the model upon which $\mathcal{L}$ is based. Those familiar with the binomial probability model will note the similarity with the binomial likelihood. The probability model of the data and the likelihood function of the parameter are closely related; they merely reverse the roles of the data and the parameters, given a model. The important point to remember is that the likelihood function is always a product of the probabilities.

Thus, given the data (y and n) and the binomial model, one can compute the *likelihood* that p is 0.15 or 0.73 or any other value between 0 and 1. The likelihood (a relative, not absolute value) is a function of only the unknown parameter p. Given this formalism, one might compute the likelihood of many values of the unknown parameter p. The likelihood of 4 values of p are tabulated below.

P	$\mathcal{L}$
0.3	0.0173
0.5	0.1611
0.7	0.2201
0.8	0.1107

Clearly, some values of p are much more *likely* than others and this is invariant to any scaling of the data. In fact, $p = 0.7$ is 12.7 (= 0.2201/0.0173) times more likely than the value of $p = 0.3$. Given the ability to compute the likelihood of various values of p, Fisher reasoned that the best estimate of the unknown parameter p would be the one that was “most likely.” Hence the term *maximum likelihood estimate* or MLE. In the values tabulated above, $p = 0.7$ is the most likely. If the derivative of the analytical form of the likelihood were used to compute the exact maximum of the entire function, we would see that the MLE is 0.63636. This estimate could also be gotten using numerical methods and that is what is done in practice with real problems.

It seems compelling to pick the value of the unknown p that is “most likely.” Likelihood theory includes asymptotically optimal methods for estimation of unknown parameters and their variance–covariance matrix, derivation of hypothesis tests, the basis for profile likelihood intervals, and other important quantities. Likelihood is the backbone of statistical theory, whereas least squares can be viewed as a limited, special case (but certainly an important case).

A.2 Log-Likelihood Functions

For many purposes the natural logarithm of the likelihood function is essential; written as log($\mathcal{L}(\theta|$data, model)), or log($\mathcal{L}$ ($\theta|$y, model)), or if the context is clear, just log($\mathcal{L}(\theta)$) or even just log($\mathcal{L}$). Thus, taking logarithms

$$\log(\mathcal{L}(\theta \mid y, \text{model})) = \log\binom{n}{y} + y \cdot \log(p) + (n-y) \cdot \log(1-p).$$

Often, one sees notation such as log($\mathcal{L}$ ($\theta|$y)), without making it clear that a particular model is assumed. An advanced feature of log($\mathcal{L}$) is that it, by itself, is a type of *information* concerning the unknown parameters (θ) and the model. A property of logarithms for values between 0 and 1 is that they lie in the negative quadrant; thus, values of the log-likelihood function are negative (unless some additive constants have been omitted). Figure A.1 shows a

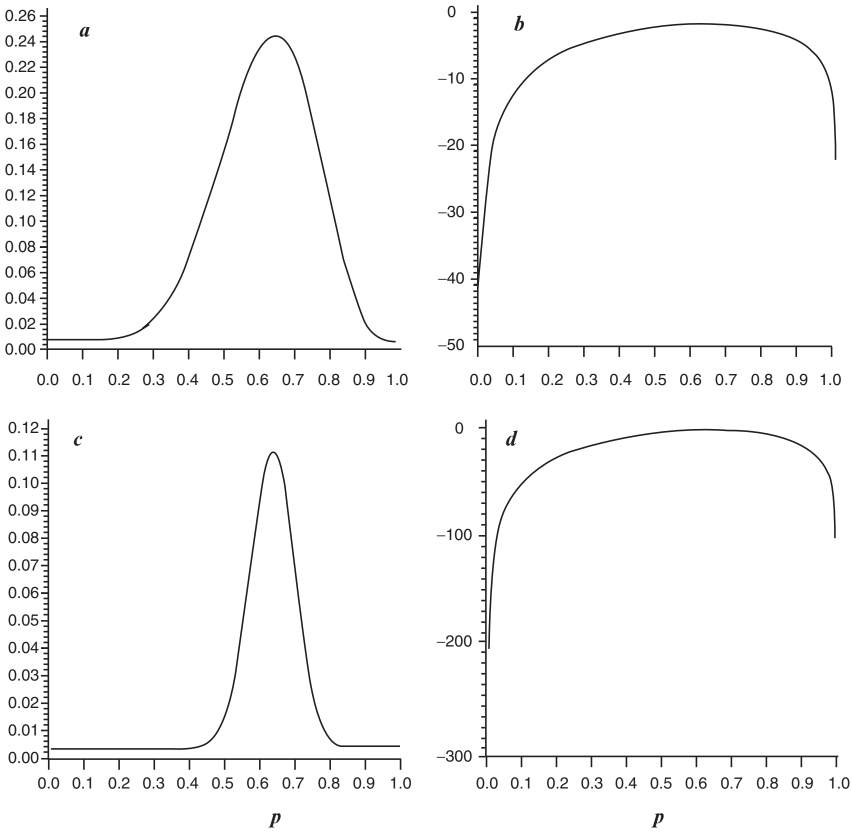

FIG. A.1. Plots of the binomial likelihood (*a*) and log-likelihood (*b*) function, given $n = 11$ penny flips and the observation that $y = 7$ of these were heads. Also shown are plots of the binomial likelihood (*c*) and log-likelihood (*d*) function, given a sample size 10 times larger; $n = 110$ penny flips and the observation that $y = 70$ of these were heads.

plot of the likelihood (*a*) and log-likelihood (*b*) functions where 11 coins are flipped, 7 heads are observed, and the binomial model is assumed. The value of $p = 0.6363$ maximizes both the likelihood and the log-likelihood function; this value is denoted as $\hat{p}$ and is the maximum likelihood estimate (MLE). Relatively little information is contained in such a small sample size (n = 11) and this is reflected in the broad shape of the plots. Had the sample size been 10 times larger, with $n = 110$ and 70 heads observed, the likelihood and log-likelihood functions would be more peaked (Fig. A.1). In fact, the sampling variance is derived from the shape of the log-likelihood function around its maximum point. Finally, the value of the log-likelihood function at its maximum point is a very important quantity and it is this point that defines the maximum likelihood estimate. In the example with 11 flips and 7 heads, the value of the maximized log-likelihood is −1.411 (Fig. A.1b). Thus, when one sees reference to a maximized $\log(\mathcal{L}(\theta))$, this merely represents a numerical value (e.g., −1.411). The value −1.411 is computed using the binomial coefficient

$$\binom{11}{7} = \frac{11!}{7!(11-7)!} = 330.$$

Specifically, the value of the maximized log-likelihood function is

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

$$\begin{aligned}\log(\mathcal{L}\,(p \mid 7{,}11{,}\text{binomial})) &= \log\binom{11}{7} + 7 \cdot \log(0.6363) + (4) \cdot \log(1-0.6363)\\ &= 5.799 + 7(-0.452) + 4(-1.012)\\ &= -1.411.\end{aligned}$$

The value of the log-likelihood function $\log(\mathcal{L}) = -1.411$. Then, AIC = $-2\log(\mathcal{L}) + 2K$ is simply $-2(-1.411) + 2(1) = 4.822$. Software for computing MLEs always give the value of the maximized log-likelihood or the deviance (which is $-2\log(\mathcal{L})$ and is the first term in AIC and AICc). Thus, computation of AICc is trivial once the MLEs have been found.

Those using LS to get estimates in linear models can easily compute the value of the maximized log-likelihood function by the simple mapping

$$\log\left(\mathcal{L}\,(\hat{\underline{\theta}})\right) \sim -\tfrac{1}{2} n \log(\hat{\sigma}^2),$$

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

The variance–covariance matrix can be found from the log-likelihood function; this is a more technical subject and I will only provide a glimpse into Fisher's approach. The variance is directly related to the shape (peakedness) of the log-likelihood function near the maximum point. The more peaked the smaller the variance and vice versa. If there are 3 unknown parameters, then the variance–covariance is a square matrix with dimension 3. The 3 variances appear on the diagonal, while the covariances appear in the off-diagonal elements. [Elements of this matrix come from second mixed partial derivatives of the log-likelihood function with respect to the parameters. This is a very general and useful procedure, but often seems difficult when first encountered; we will not take this issue further here.]

The likelihood function $\mathcal{L}(\theta|x$, model) makes it clear that for inference about θ the data and the model are taken as *given*. Before one can compute the likelihood that $\theta = 0.53$, one must have data and a particular statistical model. While an investigator will have empirical data for analysis, it is unusual that the model is known or given. Rather, a number of alternative model forms must be considered as well as the specific explanatory variables to be used in modeling a response variable. This issue includes the *variable selection problem* in multiple regression analysis. If one has data and a model, LS or ML theory can be used to estimate the unknown parameters (θ) and other quantities useful in making statistical inferences.

Model selection relates to fitted models; given the data and the form of the model, then the MLEs of the model parameters have been found ("fitted").

A.3 Why Likelihood Theory?

The review above has been in terms of only one model (the binomial) with a single unknown parameter, but the principles extend to other models and models with hundreds of unknown parameters. The theory is worth the effort to learn and be comfortable with. Reasons for this include

- Likelihood and log-likelihood functions form the general basis for deriving estimates of unknown parameters in the models of science hypotheses and their variance–covariance matrix as measures of precision
- Log-likelihood functions are the basis for profile likelihood intervals. These allow for asymmetric intervals and avoid the notion of repeated sampling and the awkward definition of the usual frequentist intervals
- Likelihood and log-likelihood values are the basis for hypothesis tests – the likelihood ratio tests (LRT) and goodness-of-fit tests in particular (however, these are of little use in model building or model selection)
- Model selection based on Kullback-Leibler information

A.4 Properties of Maximum Likelihood Estimators

MLEs are asymptotically optimal; that is, as sample size gets "large" they enjoy the following important properties:

- Normally distributed
- Minimum variance
- Unbiased

In addition, linear or nonlinear transformations of an MLE to estimate another parameter are also MLE. For example, mean life span $\bar{L}$ is defined as 1/log(S). An estimator of mean life span is

$$\hat{\bar{L}} = 1/\log(\hat{S}),$$

where $\hat{S}$ in an MLE. This being the case, then one can say that $\hat{\bar{L}}$ is also MLE. This is a very important property in application.

A.5 Deviance

A useful quantity in likelihood-based inference is the deviance,

$$\text{Deviance} = -2\log(\mathcal{L}(\hat{\theta} \mid x{,}g)) + 2\log(\mathcal{L}_{\text{s}}(\hat{\theta} \mid x, g)),$$

where $\mathcal{L}$ is a "saturated" model. In model selection, this $\mathcal{L}_{\text{s}}$ term is constant across models and can usually be omitted. In other situations the saturated model would produce $\log(\mathcal{L}_{\text{s}}) = 0$; hence, there is a basis to say deviance = $-2\log(\mathcal{L}(\hat{\theta}|x, g))$. Thus, for the issues here, deviance = $-2\log(\mathcal{L}(\hat{\theta}|x, g))$ and is a measure of lack of fit and is the first term in AIC and AICc.

A.6 Likelihood Ratio Tests

Likelihood ratio tests (LRT) can be used to compare two nested models; the form of the test is suggested by its name

$$T = -2\log\left(\frac{\mathcal{L}_s(\hat{\theta}\,|\,x{,}g)}{\mathcal{L}_g(\hat{\theta}\,|\,x{,}g)}\right),$$

where the simpler model (s) has fewer parameters than the general model (g) – seen as subscripts. [Note the appearance of the −2 again.]

Asymptotically, the test statistic (T) is distributed as a chi-squared variable with degrees of freedom equal to the difference in the number of parameters between the two nested models. LRTs can also be expressed in terms of the difference between the two deviances.

A.7 A Likelihood Version of R^2

Nagelkerkle (1991) provided a near analog to the R^2 of least squares, we will denote this as $\mathcal{R}^2$. Let $\ell(\hat{\theta})$ and $\ell(0)$ denote the maximized log-likelihoods for the fitted model of interest and the null model, respectively. Start with

$$R^2 = 1 - \exp\left\{-\frac{2}{n}(\ell(\hat{\theta}) - \ell(0))\right\}$$

and then rescale to allow a maximum of 1 by defining

$$\max R^2 = 1 - \exp\left\{\frac{2}{n}\cdot \ell(0)\right\}.$$

and finally the rescaled value we want

$$\mathcal{R}^2 = R^2 / \max R^2.$$

Often, the statistic $\mathcal{R}^2$ is optimistic and it is not an exact analog to the usual R^2 in linear models. Still, this approach is useful and easy to compute. In addition, other approaches have been developed such as the "analysis of deviance," which is closely allied with the usual R^2 in regression.

A.8 Potential Problems

Virtually all applications of likelihood methods for real problems are done numerically. That is, calculus is not used to find the maximum of multidimensional functions; instead, sophisticated numerical routines have been found years ago to perform these tasks.

The first problem is that the function, at least in one dimension, is very flat and the numerical routine cannot identify the "exact" maximum point. There are several reasons that might cause this; however, the software usually outputs a message that it failed to converge. The user might restart the routine using the provisional values of $\hat{\theta}$ available when the routine last stopped. Alternatively, one might start over using a different starting value for $\hat{\theta}$.

The second problem is that a log-likelihood function might have multiple local maxima (modes) and one must worry that the numerical routine will find a suboptimal maximum point and this is unknown to the user. Here, one might try different starting values or use some other numerical routine (e.g., simulated annealing). Most of the commonly used statistical distributions are in the so-called "exponential family" and these carry a guarantee of unimodality (however, mixture distributions of these common forms do not).

Appendix B: Expected Values

Statistical expectations of estimators or other expressions are often useful in a variety of ways. Such expectations can be thought of as an "average" taken over all possible samples of size *n* (see Wackerly and Mendenhall 1996). This process is simple when working with discrete random variables. The expectation of a discrete random variable *x* is defined as

$$\mathbf{E}(x) = \sum_{\mathrm{i}} x_i p(x_i),$$

where *p* is the probability of being in class i. Consider a population of size N = 4 and a sample of size 2. The binomial coefficient $\binom{N}{n}$ is read "N chose n" or, in this example, $\binom{4}{2}$ is "4 chose 2" = 4!/[2! × (4–2)!] = 6. This is an effective way to compute the number of ways a sample of size 2 can be drawn from a population of size 4. In general, the binomial coefficient is

$$\binom{N}{n} = \frac{N!}{n! \times (N-n)!},$$

where ! means factorial. Let $N = 5$, then 5! is $5 \times 4 \times 3 \times 2 \times 1 = 120$.
Now consider a population of 4 rats (rat A, B, C, and D) each with a number of ticks. An exact count of the number of ticks on each rat has been made; rat A has 2 ticks, rat B has 4 ticks, rat C has 2 ticks, and rat D has 8 ticks. As we have an exact count of the number of ticks on all the rats in the population, we can compute the mean number of ticks per rat as a population parameter; denote this parameter as μ. The value of μ in this simple example is merely the total number of ticks (2 + 4 + 2 + 8 = 16) divided by the number of rats (4). This gives the parameter as μ = an average of 4 ticks per rat. So, the population parameter in this example is known, $\mu = 4$.
We must now summarize all possible samples of size 2 that could be drawn from the population of size 4; we know from the binomial coefficient that there are 6 such samples of size 2 possible. The sample data are summarized below:

Sample, *i*	No. ticks	Sample mean	$\hat{\mu}$
1	AB	6	3
2	AC	4	2
3	AD	10	5
4	BC	6	3
5	BD	12	6
6	CD	10	5

Each of the 6 sample means $\hat{\mu}_i$ is a maximum likelihood estimate. The expected value of the MLE $\hat{\mu}$ is written as E($\hat{\mu}$) and is the average of the 6 sample means

$$(3+2+5+3+6+5)/6=4.$$

Thus, $E(\hat{\mu})$ = 4. This is the average of all possible samples from the population of size 4 for samples of size 2. The notation "$E(\bullet)$" is an operator meaning "take the expectation of (•)." One reason for taking statistical expectations is in assessing the bias of an estimator. Bias is also an average quantity and defined as

$$\text{Bias} = E(\hat{\theta}) - \theta$$

where θ is some parameter of interest. In the rat example, bias = $\mathrm{E}(\hat{\mu}) - \mu = 4 - 4 = 0$, or unbiased. Expectations of continuous random variables also exist; integrals replace summation operators, but the principle remains the same.

A second type of expectation is useful in parameterizing some types of models. Consider a sample of size R_2 sea turtles marked and released in year 2 of a conservation biology study. Four years after release, r_{25} turtles are killed (as bycatch) in a primitive fishery and reported to conservation authorities. The notation r_{25} reflects the number of turtles recovered dead in year 5 from those marked and released in year 2. So, under a model that allows survival and reporting probabilities to vary by year, we can write down the expectation of r_{25}, i.e., $E(r_{25})$. Here the expectation operator (E) asks for the analytical expression of the count r_{25}, given a model. We note that to have been killed and reported in year 5, the turtles had to survive the yearly intervals 2–3, 3–4, 4–5, they died in year 5, and were reported in year 5. Thus, under the time-specific model

$$E(r_{25}) = R_2 S_2 S_3 S_4 (1-S_5)\lambda_5,$$

where S is the annual survival probability in year j and λ is the annual reporting probability in year j. In this case, one would like estimates of the 5 model parameters and their sampling covariance matrix using maximum likelihood methods. The expectation changes if a different model is hypothesized where the parameters are nearly constant across years (an approximation as we know that there is some variation in the parameters across years). Here

$$E(r_{25}) = R_2 SSS(1-S)\lambda = R_2 S^3(1-S)\lambda.$$

Under this model there are only 2 parameters, S and λ. The expectation operator is used often in statistics.

A final example is the expectation of an encounter history matrix used in capture–recapture and occupancy models. For each sampling occasion i let "1" denote encountered and "0" denote not encountered. As an example, take the encounter history for manatee no. 17 over 8 sampling occasions:

$$\{11001101\}.$$

The "1" in the final column (representing year 8) makes it clear that the animal was still alive in the 7th (last) year. Thus, the expectation must contain 7 annual survival probabilities, ϕ_1, ϕ_2,…, ϕ_7, related to the 7 intervals defined by the 8 occasions (this reasoning assumes the model has year-specific parameters). These models condition on the first occasion and so there is no encounter probability (denoted as p_1) for occasion 1. Note, this manatee was encountered on occasion 2, 5, 6, and 8, following its initial capture. Thus, the expectation must contain p_2, p_5, p_6, and p_8. Finally, this animal was not encountered on occasions 3, 4, and 7 and so the expectation must include $(1–p_3)$, $(1–p_4)$ and $(1–p_7)$. In summary

$$E\{11001101\} = \phi_1\phi_2\cdots\phi_7 p_2 p_5 p_6 p_8 (1-p_3)(1-p_4)(1-p_7);$$

however, the order of the parameters is arbitrary. This component of the model has 14 unknown parameters.

As above, the expectation would be different if a different model were hypothesized. For example, if one hypothesized a fairly constant environment and relatively constant sampling effort, then a model with only an average annual survival and encounter probability would yield the following expectation for the same encounter history

$$E\{11001101\} = \phi^7 p^4 (1-p)^3.$$

This model has only 2 parameters and these parameters and their covariance matrix can be estimated using maximum likelihoods methods, given data. Given a specific data set, which of these 2 models is "better"? This is a model selection problem and its solution must take into account the concept of parsimony.

Appendix C: Null Hypothesis Testing

The central inferential issues in science are twofold. First, scientists are fundamentally interested in estimates of the magnitude of parameters or functions of parameters and their precision: are the effects trivial, small, medium, large, or extra large? Are these effects biologically meaningful or interesting? This is an *estimation* problem whether the data arise from a strict experiment or an observational study. Second, one often needs to know if the effects are large enough, given the data, to justify inclusion in a model to be used for further inference (such as prediction). This is a *model selection* problem and involves the principle of parsimony. These issues are not strongly associated with null hypothesis testing, *P*-values, and rulings about "statistical significance." Null hypothesis testing in the statistical sciences is like protoplasm in biology; they both served an early purpose but are no longer very useful.

Some people still believe that statistics and statistical science are mostly about testing null hypotheses without realizing the uninteresting or trival nature of most such hypotheses. Many null hypotheses are merely strawmen to be struck down and rejected, but little understanding is gained by doing so. We need to move on from the traditional testing approach because it is so uninformative.

Given that many of us were trained in null hypothesis testing, it is easy to cling to the incorrect notion that *P*-values represent a strength of evidence. Royall (1997), Vieland and Hodge (1998:285), and Johnson (1999) provide convincing proof that this is not the case (the reasons are technical in that *P*-values are dependent upon the sample space of both observed and unobserved data). One unsettling issue (there are many) is assigning probabilities to events that were never observed. I urge people to think hard about the differences in approach as illustrated by the European dipper example in Sect. 4.8.

Some authors still see a use for null hypothesis testing when the evidence against this seems, to me, so overwhelming (e.g., Stephens et al. 2005; Steidl 2007); I do not mean to criticize, only to note the large variance component here. I believe that null hypothesis testing will continue to decline as it is replaced by the substantially more relevant methods based on information theory and Bayes' theorem.

Appendix D: Bayesian Approaches

This appendix assumes the reader has a basic understanding of the Bayesian paradigm. Bayesian approaches have seen tremendous growth and recognition in the past 2–3 decades (Gelman et al. (2003) lists nearly 600 references). This change has been the result of huge increases in computing power and the discovery of powerful numerical methods (i.e., Markov Chain Monte Carlo methods, MCMC, see Chen et al. (2000) and Givens and Hoeting (2005)). Bayesian methods are particularly powerful in coping with a wide class of random effects models (see Sect. 6.5) and will continue to see heavy use in this area. There are many excellent books on Bayesian methods including Carlin and Louis (2001) and Gelman et al. (2003).

Bayesian methods have met with controversy over the past 2.5 centuries; this stems primarily from the subjective nature of early Bayesian approaches. Change has emerged in the thinking of many Bayesians because of the use of "vague" priors; also termed uninformative, colorless, or flat priors. Here the goal is to attempt to withhold any subjective (or "personal") information; thus, the resulting analysis is objective and the parameter estimates are often virtually identical to the MLEs. This change in approach has greatly lessened the strong objection to Bayesian methods in science where subjectivity is to be minimized, not invited or enhanced. Subjective priors on parameters often have utility in nonscience issues; but such priors have been largely rejected in scientific work. Having said that, I must note that the data "swamp" the prior in some science applications and, if this is clearly demonstrated to be the case, then there are no objections with this approach in scientific work. The use of vague priors on model parameters has been a major step forward for the acceptance of Bayesian approaches by scientists.

Bayesian approaches to model selection include the Bayesian Information Criterion (BIC), the deviance information criterion (DIC), and a reverse jump Markov Chain Monte Carlo approach (RJMCMC). DIC is a Bayesian approach but with AICc-like properties and has seen heavy use in the free software WINBUGS and more generally. DIC seems to be the workhorse for Bayesian model selection; however, other approaches also see application.

Bayesian prior probabilities on models are required when dealing with several models. BIC (see Appendix E) has both a Bayesian derivation and a "frequentist" derivation, whereas AIC also has both a Bayesian and "frequentist" derivation. Thus, debate should not be just "Bayesian vs. non-Bayesian" (see Link and Barker 2006); the issues are more substantive than this. Turning beliefs about models into probability distributions has been difficult. Still, I think a goal in Bayesian analysis would be to have the model priors swamped by the data.

The level of education and experience needed to thoughtfully use Bayesian methods is fairly high. One must have a decent background in probability, mathematical statistics, numerical analysis, and programming (*R* being especially useful) in addition to the subject matter science. This is asking a lot. I encourage research people in the life sciences to seek a PhD level statistician with expertise in Bayesian theory and computation and work collaboratively with them.

Programs such as WINBUGS are useful for smaller problems and can be surprisingly useful for many research problems. Otherwise, the researcher must often write and debug code for the MCMC or RJMCMC algorithms and this can be quite challenging. One must anticipate substantial computer run times as well as programming and debugging issues. The recent text by Givens and Hoeting (2005) provides a review of these issues.

I have a high regard for Bayesian approaches and I expect to see their increasing use in the future. In multilevel random effects models, there is little choice of method and the nature of the MCMC algorithm makes Bayesian approaches a natural for coping with random effects (however, the concept of h-likelihood might provide an alternative at some point). I think more work needs to be done to explore the mutualities between extended likelihood theory and Bayesian methods. Ken Burnham has shown several areas of commonality between what might be called likelihoodists and Bayesians (Burnham and Anderson 2004). Other investigators have found similar convergence and I view these as constructive.

Appendix E: The Bayesian Information Criterion

Akaike's AIC started one of Claude Shannon's "bandwagons," the first and best known is BIC, the Bayesian information criterion (also called SIC after its founder, Schwarz (1978)). BIC is superficially similar to AIC

$$\mathrm{BIC} = -2\log(\mathcal{L}(\hat{\underline{\theta}})) + K\log(\mathrm{n})$$

but with a different "penalty" term. If n = about 8, then BIC = AIC. In the realistic cases where $n > 8$, the penalty in BIC is slightly larger and there is a tendency for it to select smaller dimensioned models than AIC. Comparisons between BIC and AICc are harder to generalize.

BIC has nothing linking it to information theory, a misnomer. Many Bayesians do not like BIC (e.g., Link and Barker 2006); however, it is not uncommon to see its output by various statistical software packages, thus I will offer a few comments and a comparison. Almost any short summary as to what BIC is supposed to do is probably somewhat wrong or incomplete (including this one). There are a large number of papers about BIC; useful (but inconsistent) summaries can be found in Weakliem (2004). McQuarrie and Tsai (1998) provide the results of elaborate MC simulation studies that include BIC as one criterion. BIC has been rediscovered many times and several elaborations have been published over the years. I will not attempt a thorough review; instead I will offer some overview comments on this issue.

E.1 Schwarz' Criterion

Schwarz' derivation of BIC does not assume that a true model exists; however, the general setting is that a true model exists, this model is in the candidate set, and the investigator does not know which model is the true one, thus a model selection problem – "find the true model." Schwarz derived the criterion using vague priors on all the model parameters and uniform priors ($1/R$) on models. Bozdogan (1987) termed what would eventually become a class of such criteria, "dimension consistent."

Consistency is a statistical property in estimation theory indicating an estimator with both bias and variance going asymptotically to zero. Consistency has often been touted as BIC's virtue; however, this has no meaning without the false concept of a true model being in the candidate set.

E.2 Real World Properties

The real issue, then, concerns the properties of BIC when the true model is not in the set and when sample size is less than very large. Such properties are difficult to state clearly as they depend substantially on the nature of the underlying reality. I will outline two extremes, (a) are there only 3–4 large effects (and no other effects) in the underlying process? or (b) are there a wide range (say, 25–80 – if not hundreds) of tapering effect sizes in the underlying process of interest? Some useful generalizations can be given for these cases.

In (a) BIC will often do well in terms of selecting the model with these few and large effects even if sample size is small to moderate (so will AICc). Nearly all MC simulation studies generate data from a model with a few (2–5) large effects (but see McQuarrie and Tsai 1998); thus, the result would seem to show that BIC selects the true model a high percentage of the time.

In (b) BIC will perform poorly in identifying the full extent of reality unless sample size is very, very large. BIC approaches the true model from the left; thus, if sample size is too small, an underfitted model (as judged by full reality) will be selected. BIC will do poorly at selecting the model of complex reality in case (b), unless one has samples sizes in the (I am guessing) millions. Understanding the underlying realities gives little place for BIC to contribute. Link and Barker (2006) and offer additional points.

Burnham and Anderson (2002) suggested the notion of a quasi-true model to help with an understanding of BIC's performance in realistic situations; however, even this notion is strained, but at least it points to the target model for BIC selection when a true model is not in the set. BIC does not guarantee a good parsimonious model, or minimum MSE, good confidence interval coverage, or other performance properties.

E.3 High Probability Assigned to Models that Do Not Fit

BIC has a tendency to give high weight to models that do not fit, as judged by a usual goodness-of-fit test (Burnham and Anderson 2004:293–297). One might hope that if the global model fits, the selected model would also fit: AICc has this property. Under tapering effect sizes and using $\alpha = 0.05$, they found that BIC selected nonfitting models 11.5% of the time with sample size = 50, 15.9% of the time with sample size = 100, and 28.1% of the time with sample size of 500. As sample size increases, the probability of selecting a nonfitting model increases! These results would seem to be disturbing and more work on this issue is warranted.

Reschenhofer (1996) noted that AICc and BIC have very different objectives and target models and should not be directly compared. AICc depends on the given sample size and selects the *fitted* model that minimizes estimated, expected K–L information as the approximating model of full reality. AICc is about approximation and prediction and its target model changes with changes in sample size. Thus, as sample size gets larger, additional effects can be uncovered; this includes reality where there are countless tapering effects. AICc is about "best" models in the sense of approximations to truth and out-of-sample prediction, given the sample size.

E.4 Predictive Mean-Squared Error

Almost no MC simulation studies have been reported in the literature where data were generated from a model with reasonable complexity (say, a nonlinear model with 40–50 or 100 parameters, many correlated covariates, several higher order interactions). Then, over a range of sample sizes, evaluate various selection criteria on predictive mean-squared error (PMSE) or achieved confidence interval coverage for predictions. Burnham and Anderson (2002:300) present the results of a reanalysis of the human body fat data from Johnson (1996). This is a linear regression to predict body fat using 13 predictor variables (= 8,191 models). They took the global model, its MLEs, and covariance matrix and used it as a generating model to simulate 10,000 reps each with sample size 252. I will not give details here except to tabulate some PMSEs ($\times 10^6$) for (a) model averaging (multimodel inference, Chap. 5) used or (b) inference from (only) the best model.

Method	Model Averaged	Best Model
AICc	4.8534	5.6849
BIC	5.8819	7.6590

AICc has a substantially better PMSE, but note that BIC benefited relatively more from model averaging. More simulation studies to mimic real world phenomenon would be helpful. In these cases, the evaluation should be focused on PMSE instead of the usual "how often does this criterion select the true model"? Of course, the generating ("true") model should not be in the set, a mistake so often seen in the literature.

In summary I would not use BIC unless I was trying to select the generating model from MC simulation. There, a true (generating) model exists and I know if it is in the set. Then, if the generating model mimicked some complex reality and if sample size is very large (e.g., perhaps hundreds of thousands or millions), I would use BIC. Alternatively, if I knew the underlying process had 3–5 large effects (and no smaller effects) I might use BIC even if sample size was modest – this is BIC's element. Putting this in perspective, I would still use BIC in regression settings over step-up, or step-down or stepwise methods in regression.

Appendix F: Common Misuses and Misinterpretations

The recent literature in a cross section of the life sciences suggests several problem areas. I will explain a dozen of these including my own observations along with ideas suggested by various reviewers. Related suggestions are found in Anderson and Burnham (2002). Some other comments and opinions are given at www.warnercnr.colostate.edu/~anderson/PDF_files/AIC%20Myths%20and%20Misunderstandings.pdf

1. Often too little time is devoted to generating a good set of alternative hypotheses. Some published papers seem to suggest that this important step was almost an afterthought. It might be useful for an investigator preparing to collect data to ask himself “how much effort was put into developing my specific objectives and outlining the alternative hypotheses.” If the answer is “a few hours,” then it might be best to revisit these important issues.
2. Some authors tend to ignore sample size issues when interpreting model selection results and then compounding this by misinterpreting the results in a dichotomous yes/no fashion (e.g., “... uptake rates did not vary across study groups” or “...there was no difference in transition probability by group”). Of course, rates of uptake and transition probabilities differed; the issue is “by how much”? Perhaps they meant to say that with the sample size available, differences in uptake seemed small. Or perhaps, the estimated differences were large, but the sample size was so small that models with such differences could not be supported.

   The lowest level of reliable inference is the sign of the effect (+ or −). If even the evidence for the sign is weak, perhaps judgment should be withheld. The parameter estimate and its confidence interval could be given but one should probably admit that the estimated effect is about 0 as far as the information in the data are concerned.
3. Some papers misinterpret the relative importance of models within about $2\Delta_i$ units when there is no change in the deviance and differing by only one parameter (the “pretending variable problem”). This issue is aggravated when the values of the maximized log-likelihoods or the deviance are not tabled (see Anderson et al. 2001b).
4. Other literature has appeared where model building, fitting, selection, and inference are treated piecemeal (e.g., splitting a dataset for purposes of “validation,” treating groups, such as gender, separately without hypothesizing that some parameters might be in common across groups, ignoring the principle of parsimony in hypothesizing and modeling). These are not easy issues to understand but sometimes available software will help with this issue (e.g., the R package of freeware, Venables and Smith (2002)).
5. Some papers provide only a table of AICc and Δ_i values allowing a ranking of the models and their hypotheses. This approach might have been reasonable 10–15 years ago; however, much more can be learned using the model probabilities, evidence ratios, and model-averaged parameter estimates to gain insights into estimated effect sizes and structural relationships. In any case, inference should not stop at just identifying the “best model” as estimates of model parameters should be interpreted and these insights should be tied back to the science hypotheses.
6. A large percentage of papers present the results of simple studies as a NHT (see Stephens et al. 2005); perhaps without realizing that an evidence ratio would be easier to compute and provide a proper strength of evidence for *both* the null and the alternative (e.g., the model probabilities). The information-theoretic approach provides the probability of both the null and the

probability of the alternative, model-averaged estimates of effect size and estimates of precision that include a variance component for model selection uncertainty (see Schmidt et al. (2004) for a nice example).

7. Perhaps the worst issue is the feeling among some people that the information-theoretic methods "require" sustained thinking leading to hypothesizing good alternatives and this is too hard and too much to expect (see discussion by Steidl (2007)). Therefore, the NHT approach might be preferred because less thinking is required (i.e., one can always trump up a null). This attitude often fosters people spending resources playing the "measuring nature game" without much purpose. Alternative science hypotheses and hard thinking represent the very core of good science; good science is not always "easy." I doubt if anyone has received a Nobel Prize for testing a null hypothesis.
8. I often hear that some authors are encouraged/forced by journal editors or associate editors to add *P*-values in place of (or in addition to) estimates of effect size and their confidence intervals and model selection statistics. It seems, to me, that the peer review process could be much better. A colleague suggested that the weakest link in our science is that the accumulation of supposed knowledge is based on the unsupervised individual application of statistical hypothesis testing with very little effective oversight in the review process.
9. I see where investigators have conducted all-possible paired comparisons using *t* tests and then used those that were statistically "significant" from the null in a multiple regression model (i.e., the "nonsignificant" variables are discarded). This is often followed by discarding the variables in the regression model that are then not "significant." This procedure attempts to "weed out" nonsignificant variables before moving to a multivariable regression analysis with a further weeding of those found to be nonsignificant once the "more comprehensive" modeling started. This strategy is not without its logic if one has no background in statistical theory and stochasticities.

   However, this strategy is very poor for several important reasons (e.g., it mixes analysis paradigms, leads to a host of technical matters such as the multiple testing problem, and often makes hidden assumptions concerning independence of the predictor variables). If the simple models represented plausible hypotheses, they should have been in the candidate set of the initial regression models. Underlying this type of error is that the focus of the investigation has improperly focused on models rather than concentrating on the science issues (i.e., plausible hypotheses). The situation points to poor study design that often stems from shallow thinking about the science issue in the first place. This problematic approach is rampant in some areas of the life sciences.
10. The use of too many models is problematic. This is often the result of a focus on running models versus thinking about the alternative science hypotheses. Certainly, if there are more models than the size of the sample

($R > n$), one should expect difficulties. Large, unfocused descriptive studies are often faced with a huge number of models (see no. 11, below).

11. Some software packages allow one to perform a “stepwise AIC” and this represents poor practice. The theme here seems to be that the computer will “find out what is important without the investigator having to think.” The underlying problem, like running “all possible models,” is the finding of effects that are, in fact, spurious. This issue relates back to Freedman’s paradox and model selection bias. Admittedly, these are issues that are not easy to understand without some background.
12. In general, I think the results from rampant data dredging should often remain unpublished. I think more should be done to explain to readers what results and conclusions stem from a priori considerations versus the more tentative insights from *post hoc* investigations. Such statements portray honesty and openness in publication and can help define the next set of hypotheses and their models.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. *in* B. N. Petrov, and F. Csaki (Eds.) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest. pp. 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* **19**, 716–723.

Akaike, H. (1977). On entropy maximization principle. *in* P. R. Krishnaiah (Ed.) *Applications of statistics*. North-Holland, Amsterdam. pp. 27–41.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* **30**, 9–14.

Akaike, H. (1981a). Likelihood of a model and information criteria. *Journal of Econometrics* **16**, 3–14.

Akaike, H. (1981b). Modern development of statistical methods. *in* P. Eykhoff (Ed.) *Trends and progress in system identification*. Pergamon Press, Paris. pp. 169–184.

Akaike, H. (1983a). Statistical inference and measurement of entropy. *in* G. E. P. Box, T. Leonard, and C-F. Wu (Eds.) *Scientific inference, data analysis, and robustness*. Academic Press, London. pp. 165–189.

Akaike, H. (1983b). Information measures and model selection. *International Statistical Institute* **44**, 277–291.

Akaike, H. (1985). Prediction and entropy. *in* A. C. Atkinson and S. E. Fienberg (Eds.) *A celebration of statistics*. Springer, New York, NY. pp. 1–24.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* **52**, 317–332.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. *in* S. Kotz, and N. L. Johnson (Eds.) *Breakthroughs in statistics, Vol. 1*. Springer-Verlag, London. pp. 610–624.

Akaike, H. (1994). Implications of the informational point of view on the development of statistical science. *in* H. Bozdogan, (Ed.) *Engineering and Scientific Applications, Vol. 3*, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer, Dordrecht, Netherlands. pp. 27–38.

Anderson, D. R. (2001). The need to get the basics right in wildlife field studies. *Wildlife Society Bulletin* **29**, 1294–1297.

Anderson, D. R., and Burnham, K. P. (1999). General strategies for the collection and analysis of ringing data. *Bird Study* **46** (Supplement), S261–S270.

Anderson, D. R., and Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management* **66**, 912–918.

Anderson, D. R., Burnham, K. P., Franklin, A. B., Gutierrez, R. J., Forsman, E. D., Anthony, R. G., White, G. C., and Shenk, T. M. (1999). A protocol for conflict resolution in analyzing empirical data related to natural resources controversies. *Wildlife Society Bulletin* **27**, 1050–1058.

Anderson, D. R., Burnham, K. P., Gould, W. R., and Cherry, S. (2001a). Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin* **29**, 311–316.

Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* **64**, 912–923.

Anderson, D. R., Burnham, K. P., and White, G. C. (1994). AIC model selection in overdispersed capture–recapture data. *Ecology* **75**, 1780–1793.

Anderson, D. R., Burnham, K. P., and White, G. C. (1998). Comparison of AIC and CAIC for model selection and statistical inference from capture–recapture studies. *Journal of Applied Statistics* **25**, 263–282.

Anderson, D. R., Burnham, K. P., and White, G. C. (2001). Kullback-Leibler information in resolving natural resource conflicts when definitive data exist. *Wildlife Society Bulletin* **29**, 1260–1270.

Anderson, D. R., Link, W. A., Johnson, D. K., and Burnham, K. P. (2001b). Suggestions for presenting the results of data analysis. *Journal of Wildlife Management* **65**, 373–378.

Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W., and Weisberg, H. I. (1980). *Statistical methods for comparative studies*. John Wiley, New York, NY.

Anthony, R. G., Forsman, E. D., Franklin, A. B., Anderson, D. R., Burnham, K. P., White, G. C., Schwarz, C. J., Nichols, J. D., Hines, J. E., Olson, G. S., Ackers, S. H., Andrews, L. S., Biswell, B. L., Carlson, P. C., Diller, L. V., Dugger, K. M., Fehring, K. E., Fleming, T. L., Gerhardt, R. P., Gremel, S. A., Gutierrez, R. J., Harpe, P. J., Herter, D. R., and Higley, J. M. (2006). Status and trends in demography of Northern Spotted Owls, 1985–2003. *Wildlife Monograph* **163**, 1–48.

Atilgan, T. (1996). Selection of dimension and basis for density estimation and selection of dimension, basis and error distribution for regression. *Communications in Statistics – Theory and Methods* **25**, 1–28.

Atmar, W. 2001. A profoundly repeated pattern. *Bulletin of the Ecological Society of America* **26**, 208–211.

Azzalini, A. (1996). *Statistical inference based on the likelihood*. Chapman and Hall, London, UK.

Ball, L. C., Doherty, P. F., Jr., and McDonald, M. W. (2005). An occupancy modeling approach to evaluating Palm Springs ground squirrel habitat model. *Journal of Wildlife Management* **69**, 894–904.

Bedrick, E. J., and Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics* **50**, 226–231.

Blanckenhorn, W. U., Hellriegel, B., Hosken, D. J., Jann, P., Altweg, R., and Ward, P. I. (2004). Does testis size track expected mating success in yellow dung flies? *Functional Ecology* **18**, 414–418.

Boltzmann, L. (1877). Uber die Beziehung zwischen dem Hauptsatze der mechanischen Warmetheorie und der Wahrscheinlicjkeitsrechnung respective den Satzen uber das Warmegleichgewicht. *Wiener Berichte* **76**, 373–435.

Bortz, D. M., and Nelson, P. W. (2006). Model selection and mixed-effects modeling of HIV infection dynamics. *Bulletin of Mathematical Biology* **68**, 2005–2025.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association* **71**, 791–799.

Box, G. E. P. (1979). Robustness in scientific model building. *in* R. L. Launer and G. N. Wilkinson, eds., *Robustness in statistics*. Academic Press, New York, NY. pp. 201–236.

Box, G. E. P., Leonard, T., and Wu, C.-F. (Eds.) (1981). *Scientific inference, data analysis, and robustness*. Academic Press, London.

Box, J. F. (1978). *R. A. Fisher: the life of a scientist*. John Wiley, New York, NY. pp. 511.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: *X*-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

Breiman, L., and Freedman, D. F. (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association* **78**, 131–136.

Broda, E. (1983). *Ludwig Boltzmann: Man, physicist, philosopher*. (translated with L. Gay). Ox Bow Press, Woodbridge, Connecticut, USA.

Brown, D., and Rothery, P. (1993). *Models in biology: Mathematics, statistics and computing*. John Wiley, New York, NY.

Brownie, C., Anderson, D. R., Burnham, K. P., and Robson, D. S. (1985). *Statistical inference from band recovery data – a handbook*. 2nd Ed. U. S. Fish and Wildlife Service Resource Publication 156. pp. 305.

Brush, S. G. (1965). *Kinetic theory, Vol. 1*. Pergamon Press, Oxford.

Brush, S. G. (1966). *Kinetic theory, Vol. 2*. Pergamon Press, Oxford.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 603–618.

Burnham, K. P., and Anderson, D. R. (1992). Data-based selection of an appropriate biological model: the key to modern data analysis. *in* D. R. McCullough, and R. H. Barrett (Eds.) *Wildlife* 2001: *Populations*. Elsevier, London. pp. 16–30.

Burnham, K. P., and Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* **28**, 111–119.

Burnham, K. P., and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach, 2nd Ed.*, Springer-Verlag, New York, NY.

Burnham, K. P., and D. R. Anderson. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research* **33**, 261–304.

Burnham, K. P., Anderson, D. R., and White, G. C. (1994). Evaluation of the Kullback–Leibler discrepancy for model selection in open population capture-recapture models. *Biometrical Journal* **36**, 299–315.

Burnham, K. P., Anderson, D. R., and White, G. C. (1995b). Selection among open population capture–recapture models when capture probabilities are heterogeneous. *Journal of Applied Statistics* **22**, 611–624.

Burnham, K. P., Anderson, D. R., and White, G. C. (1996). Meta-analysis of vital rates of the Northern Spotted Owl. *Studies in Avian Biology* **17**, 92–101.

Burnham, K. P., Anderson, D. R., White, G. C., Brownie, C., and Pollock, K. H. (1987). *Design and analysis methods for fish survival experiments based on release-recapture*. American Fisheries Society, Monograph **5**, 437.

Burnham, K. P., White, G. C., and Anderson, D. R. (1995a). Model selection in the analysis of capture-recapture data. *Biometrics* **51**, 888–898.

Caley, P., and Hone, J. (2002). Estimating the force of infection; *Mycobacterium bovis* infection in feral ferrets *Mustela furo* in New Zealand. *Journal of Animal Ecology* **71**, 44–54.

Caley, P., and Hone, J. (2005). Assessing the host disease status of wildlife and the implications for disease control: *Mycobacterium bovis* infection in feral ferrets. *Journal of Animal Ecology* **42**, 708–719.

Carlin, B. P., and Louis, T. A. (2001). *Bayes and empirical Bayes methods for data analysis, 2nd Ed.* Chapman and Hall, New York, NY.

Carter, G. M., Stolen, E. D., and Breininger, D. R. (2006). A rapid approach to modeling species–habitat relationships. *Biological Conservation* **127**, 237–244.

Chamberlain, T. C. (1890). The method of multiple working hypotheses. *Science* **15**, 92–96. (Reprinted 1965, *Science* **148**, 754–759.)

Chatfield, C. (1991). Avoiding statistical pitfalls (with discussion). *Statistical Science* **6**, 240–268.

Chatfield, C. (1995a). *Problem solving: A statistician's guide*. Chapman and Hall, London. pp. 325.

Chatfield, C. (1995b). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* **158**, 419–466.

Chen, M.-H., Shao, Q.-m., and Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. Spring, New York, NY.

Clyde, M. (2000). Model uncertainty and health effect studies for particular matter. *Environmetrics* **11**, 745–763.

Cohen, D. (1966). Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology* **12**, 119–129.

Cohen, D. (1967). Optimizing reproduction in a randomly varying environment when a correlation may exist between the conditions at the time a choice has to be made and the subsequent outcome. *Journal of Theoretical Biology* **16**, 1–14.

Cohen, D. (1968). A general model of optimal reproduction in a randomly varying environment. *Journal of Ecology* **56**, 219–228.

Cohen, E. G. D., and Thirring, W. (Eds.) (1973). *The Boltzmann equation: Theory and applications*. Springer-Verlag, New York, NY. pp. 642.

Cohen, J., and Medley, G. (2005). *Stop working & start thinking*. Taylor & Francis Group, New York, NY.

Cook. T. D, and Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin Company, Boston, MA.

Cover, T. M., and Thomas, J. A. (1991). *Elements of information theory*. John Wiley, New York, NY. pp. 542.

Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science* **5**, 169–174.

Cox, D. R. (1995). The relation between theory and application in statistics. *Test* **4**, 207–261.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press, Cambridge, UK.

Daniel, C., and Wood, F. S. (1971). *Fitting equations to data*. Wiley-Interscience, New York, NY. pp. 342.

Dawkins, R. (1986). *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. W. W. Norton, New York, NY.

deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. *in* S. Kotz, and N. L. Johnson (Eds.) *Breakthroughs in statistics, Vol. 1*. Springer-Verlag, London. pp. 599–609.

Delury, D. B. (1954). On the assumptions underlying estimates of mobile populations. Pages 287–293 *in Statistics and Mathematics in biology*. Iowa State University, Ames.

Dijkstra, T. K. (Ed). (1988). *On model uncertainty and its statistical implications*. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York, NY. pp. 138.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

Draper, N. R., and Smith, H. (1981). *Applied regression analysis*. John Wiley, New York, NY. pp. 709.

Eberhardt, L. L., and Thomas, J. M. (1991). Designing environmental field studies. *Ecological Monographs* **61**, 53–73.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press, Cambridge, UK.

Edwards, A. W. F. (1976). *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge University Press, Cambridge. pp. 235.

Edwards, A. W. F. (1992). *Likelihood: Expanded edition*. Johns Hopkins University Press, Baltimore, Maryland.

Edwards, A. W. F. (2001). Occam's bonus. p. 128–139; in Zellner, A., Keuzenkamp, H. A., and McAleer, M. *Simplicity, inference and modelling*. Cambridge University Press, Cambridge, UK.

Elliott, L. P., and Brook, B. W. (2007). Revisiting Chamberlin (1890): Multiple working hypotheses for the 21st century. *Bioscience*, **57**, 608–614.

Eng, J. (2004). Sample size estimation: A glimpse beyond simple formulas. *Radiology* **230**, 606–612.

Everitt, B. S. (1998). *The Cambridge dictionary of statistics*. Cambridge University Press, Cambridge, UK.

Findley, D. F., and Parzen, E. (1995). A conversation with Hirotugu Akaike. *Statistical Science* **10**, 104–117.

Fisher, R. A. (1936). Uncertain inference. *Proceedings of the American Academy of Arts and Sciences* **71**, 245–258.

Flack, V. F., and Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician* **41**, 84–86.

Flather, C. H. (1992). Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeograhy* **23**, 155–168.

Ford, E. D. (2000). *Scientific method for ecological research*. Cambridge University Press, Cambridge, UK.

Forsche, B. K. (1963). Chaos in the brickyard. Science **142**, 339.

Freddy, D. J., White, G. C., Kneeland, M. C., Kahn, R. H., Unsworth, J. W., deVergie, W. J., Graham, V. K., Ellenberger, J. H., and Wagner, C. H. (2004). How many mule deer are there? Challenges of credibility in Colorado. *Wildlife Society Bulletin* **32**, 916–927.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician* **37**, 152–155.

Freedman, D. A., Navidi, W., and Peters, S. C. (1988). On the impact of variable selection in fitting regression equations. *in* T. K. Dijkstra (Ed.) *On model uncertainty and its statistical implications*. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York, NY. pp. 1–16.

Fujikoshi, Y., and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 707–716.

Gallager, R. G. (2001). Claude E. Shannon: A retrospective on his life, work, and impact. *IEEE Transactions on Information Theory* **47**, 2681–2695.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian data analysis, 2nd Ed.* Chapman and Hall, New York, NY.

Gilchrist, W. (1984). *Statistical modelling*. Chichester, Wiley and Sons, New York, NY.

Givens, G. H., and Hoeting, J. A. (2005). *Computational statistics*. John Wiley, Hoboken, NJ.

Goldman, S. (1953). *Information theory*. Constable Publishing, London, UK.

Golomb, S. W., Berlekamp, E., Cover, T. M. Gallager, R. G., Massey, J. L., and Viterbi, A. J. (2002). Claude Elwood Shannon. *Notices of American Mathematical Society* **292**, 8–16.

Golub, G. H., Health, M., and Wahba, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

Good, I. J. (1979). A. M. Turing's statistical work in World War II. *Biometrika* **66**, 393–396.

Goodman, S. N., and Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health* **78**, 1568–1574.

Gotelli, N. J., and Ellison, A. M. (2004). *A primer of ecological statistics*. Sinauer Associates, Sunderland, MA.

Greenhouse, S. W. (1994). Solomon Kullback: 1907–1994. *Institute of Mathematical Statistics Bulletin* **23**, 640–642.

Guiasu, S. (1977). *Information theory with applications*. McGraw-Hill, New York, NY.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *American Statistician* **60**, 19–26.

Hairston, N. G. (1989). *Ecological experiments: Purpose, design and execution*. Cambridge University Press, Cambridge, UK.

Hald, A. (1952). *Statistical theory with engineering applications*. John Wiley, New York, NY.

Hand, D. J. (1994). Statistical strategy: Step 1. *in* P. Cheeseman, and R. W. Oldford (Eds.) *Selecting models from data*. Springer-Verlag, New York, NY. pp. 1–9.

Hasenöhrl, F. (Ed.) (1909). *Wissenschaftiche Abhandlungen*. 3 Vols, Leipzig, Germany.

Hendry, A. P., Grant, P. R., Grant, B. R., Ford, H.A., Brewer, M. J., and Podos, J. (2006). Possible human impacts on adaptive radiation: Beak size bimodality in Darwin's finches. *Proceedings of the Royal Society*, *Series B*. Published online. 1–8.

Hilborn, R., and Mangle, M. (1997). *The ecological detective*. Princeton University Press, Princeton, NJ.

Hjorth, J. S. U. (1994). *Computer intensive statistical methods: Validation, model selection and bootstrap*. Chapman and Hall, London.

Hobbs, N. T., and Hilborn, R. (2006). Alternatives to statistical hypothesis testing in ecology: A guide to self teaching. *Ecological Applications* **16**, 5–19.

Hobson, A., and Cheng, B.-K. (1973). A comparison of the Shannon and Kullback information measures. *Journal of Statistical Physics* **7**, 301–310.

Hoeting, J. A., Madigan, D., Reftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14**, 382–417.

Hoeting, J. A., Davis, R. A., Merton, A. A., and Thompson, S. E. (2006). Model selection for geostatistical models. *Ecological Applications* **16**, 87–98.

Horner, C., and Westacott, E. (2000). *Thinking through philosophy: An introduction*. Cambridge University Press, Cambridge, UK.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B*, **60**, 271–293.

Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Hurvich, C. M., and Tsai, C-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician* **44**, 214–217.

Hurvich, C. M., and Tsai, C-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499–509.

Hurvich, C. M., and Tsai, C-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51**, 1077–1084.

Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics* **29**, 411–434.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Review* **106**, 620–630.

Jessop, A. (1995). *Informed assessments: An introduction to information, entropy and statistics*. Ellis Horwood, London. pp. 366.

Johnson, D. L. (1999). The insignificance of statistical significance. *Journal of Wildlife Management*, **63**, 763–772.

Karban, R., and Huntzinger, M. (2006). How to do ecology: a concise handbook. Princeton University Press, Princeton, NJ.

Kendall, W. L., and Gould, W. R. (2002). An appeal to undergraduate wildlife programs: Send scientists to learn statistics. *Wildlife Society Bulletin* **30**, 623–627.

Keppie, D. M. (2006). Context, emergence, and research design. *Wildlife Society Bulletin* **34**, 242–246.

Kitagawa, T. (1986). Editor's preface. *in* Y. Sakamoto, Ishiguro, M., and Kitagawa, G. *Akaike information criterion statistics*. KTK Scientific Publishing Company, Tokyo, Japan. pp. xiii–xiv.

Konishi, S., and Kitagawa, G. (2007). *Information criteria and statistical modeling*. Springer, New York, NY.

Krebs, C. J. (2000). Hypothesis testing in ecology. *in* L. Boitani and T. K. Fuller, *Research techniques in animal ecology: Controversies and consequences*. Columbia University Press, New York, NY. pp. 1–12.

Kullback, S. (1959). *Information theory and statistics*. John Wiley, New York, NY.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Kullback, S. (1987). The Kullback-Leibler distance. *The American Statistician* **41**, 340–341.

Kuhn, T. S. (1970). *The structure of scientific revolutions, 2nd Ed*. University of Chicago Press, Chicago, IL.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). *Applied linear statistical models, 4th Ed*. McGraw Hill, Chicago, IL.

Lahiri, P. (Ed.) (2001). *Model selection*. Institute of Mathematical Statistics, Lecture Note – Monograph Series **38**, 256.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. John Wiley, New York, NY.

Lebreton, J.-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecological Monograph* **62**, 67–118.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized linear models with random effects: Unified analysis via H-likelihood*. Chapman and Hall, Boca Raton, FL.

Lehmann, E. L. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science* **5**, 160–168.

Leopold, A. (1933) *Game management*. University of Wisconsin Press, Madison, WI.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist* **54**, 421–431.

Linhart, H., and Zucchini, W. (1986). *Model selection*. John Wiley, New York, NY.

Link, W. A., and Barker, R. J. (2006). Models weights and the foundations of multimodel inference. *Ecology* **87**, 2626–2635.

Lukacs, P. M., Thompson, W. L., Kendall, W. L., Gould, W. R., Doherty, P. F., Burnham, K. P., and Anderson, D. R. (2007). Concerns regarding a call for pluralism of information theory and hypothesis testing. *Journal of Animal Ecology* **44**, 456–460.

Lukacs, P. M., Burnham, K. P., and Anderson, D. R. (Ed.). Freedman's paradox: Why ecologists should worry (unpublished).

Lunneborg, C. E. (1994). *Modeling experimental and observational data*. Duxbury Press, Belmont, CA, USA. pp. 506.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **12**, 591–612.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., and Hines, J. E. (2006). *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence*. Elsevier, London, UK.

Manly, B. F. J. (1992). *The design and analysis of research studies*. Cambridge University Press, Cambridge, UK.

Massart, P. (2007). *Concentration inequalities and model selection*. Springer-Verlag, Berlin.

Mauer, B. A. (1999). *Untangling ecological complexity*. University of Chicago Press, Chicago, IL.

McCullagh, P., and Nelder, J. A. (1989). *Generalized linear models, 2nd Ed*. Chapman and Hall, New York, NY.

McCulloch, C. E. (2003). *Generalized linear models*. NSF-CBMS Regional Conference Series in Probability and Statistics **7**, 1–84.

McQuarrie, A. D. R., and Tsai, C.-L. (1998). *Regression and time series model selection*. World Scientific, London, UK.

Mead, R. (1988). *The design of experiments: Statistical principles for practical applications*. Cambridge University Press, New York, NY.

Miller, A. J. (2002). *Subset selection in regression, 2nd Ed.*, Chapman and Hall, London, UK.

Moore, B. N., and Parker, R. (1986). *Critical thinking, 5th Ed*. Mayfield Publishing Company, London, UK.

Morgan, B. J. T. (2000). *Applied stochastic modelling*. Arnold Publishing, London, UK.

Muthen, L. K., and Muthen, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling* **9**, 599–620.

Nagelkerke, N. J. K. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692.

Nelder, J. A. (1991). Generalized linear models for enzyme-kinetic data. *Biometrics* **47**, 1605–1615.

Nichols, J. D. (2001). Using models in the conduct of science and management of natural resources. *in* Shenk, T., and Franklin, A. B. (Eds.) *Modeling in natural resource* *management: Development, interpretation, and application*. Island Press, Washington, D. C. pp. 11–34.

O'Connor, R. J. (2000). Why ecology lags behind biology. *The Scientist* **14**, 35.

Oliver, J. E. (1991). *The incomplete guide to the art of discovery*. Columbia University Press, New York, NY.

Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120–125.

Pan, W. (2001b). Model selection in estimating equations. *Biometrics* **57**, 529–534.

Parzen, E. (1994). Hirotugu Akaike, statistical scientist. *in* H. Bozdogan (Ed.) *Engineering and Scientific Applications, Vol. 1*, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Dordrecht, Netherlands. pp. 25–32.

Parzen, E., Tanabe, K., and Kitagawa, G. (Eds.) (1998). *Selected papers of Hirotugu Akaike*. Springer-Verlag, New York, NY.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford Science Publications, Oxford, UK.

Peirce, C. S. (1955). Abduction and induction. *in J*. Buchler (Ed.) *Philosophical writings of Peirce*. Dover, New York, NY. pp. 150–156.

Pigliucci, M. (2002a). Are ecology and evolutionary biology "soft" sciences? *Annals of Zoological Fennici* **39**, 87–98.

Pigliucci, M. (2002b). *Denying evolution: Creationism, scientism, and the nature of science*. Sinauer Associates, Sunderland, MA.

Pistorius, P. A., Bester, M. N., Kirkman, S. P., and Boveng, P. L. (2000). Evaluation of age- and sex-dependent rates of tag loss in southern elephant seals. *Journal of Wildlife Management* **64,** 373–380.

Platt, J. R. (1964). Strong inference. *Science* **146**, 347–353.

Popper, K. R. (1959). *The logic of scientific discovery*. Harper and Row, New York, NY.

Potscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory* **7**, 163–185.

Qin, J., and Lawless, G. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics* **22**, 300–325.

Rao, C. R. (2004). Forward. *in* Taper, M. L., and Lele, S. R. *The nature of scientific evidence: Statistical, philosophical, and empirical considerations*. University of Chicago Press, Chicago, IL. pp. xi–xiii.

Remontet, L., Bossard, N., Belot, A., Esteve, J., and the French network of cancer registries. (2006). An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine* **26**, 2214–2228.

Rencher, A. C., and Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics* **22**, 49–53.

Renshaw, E. (1991). *Modelling biological populations in space and time*. Cambridge University Press, Cambridge, UK.

Reschenhofer, E. (1996). Prediction with vague prior knowledge. *Communications in Statistics – Theory and Methods* **25**, 601–608.

Resetarites, W. J., Jr., and Bernardo, J. (Eds.) (2001). *Experimental ecology: Issues and perspectives*. Oxford University Press, UK.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific, Series in Computer Science, Vol 15, Singapore.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* **42**, 40–47.

Rissanen, J. (2007). Information and complexity in statistical modeling. Springer, New York, NY.

Rosenbaum, P. R. (2002). *Observational studies, 2nd Ed.*, Springer-Verlag, New York, NY.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. Chapman and Hall, London, UK.

Romesburg, H. C. (2002). *The life of the creative spirit*. Xlibris Corporation, Rome, IT.

Sakamoto, Y. (1991). *Categorical data analysis by AIC*. KTK Scientific Publishers, Tokyo.

SAS Institute Inc. (2004). *SAS® Language guide for personal computers*, Version 9.1 Edition. SAS Institute Inc, Cary, North Carolina.

SAS Institute. (2004). SAS/STAT® user's guide. 6th Ed. SAS Institute, Cary, NC.

Scheiner, S. M., and Gurevitch, J. (Eds.) (1993). *Design and analysis of ecological experiments*. Chapman and Hall, London.

Schmidt, B. R., Feldmann, R., and Schaub, M. (2004). Demographic processes underlying population growth and decline in *Salamandra salamandra. Conservation Biology* **19**, 1149–1156.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sclove, S. L. (1987). Application of some model-selection criteria to some problems in multivariate analysis. *Psychometrika* **52**, 333–343.

Seghouane, A.-K. (2005). Multivariate model selection with KIC for extrapolated cases. *Neural Networks*, Proceedings, 2005 IEEE International Joint Conference **2**, 1292–1295.

Seghouane, A.-K. (2006). Multivariate regression model selection from small samples using Kullback's symmetric divergence. *Signal Processing* **86**, 2074–2084.

Severini, T. A. (2000). *Likelihood methods in statistics*. Oxford University Press, Oxford, UK.

Shadish, W. R., Cook, T., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company, New York, NY.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 and 623–656.

Shenk, T., and Franklin, A. B. (2001). *Modeling in natural resource management: Development, interpretation, and application*. Island Press, Washington, D. C.

Shi, R., and Tsai, C.-L. (2002). Regression model selection – a residual likelihood approach. *Journal of the Royal Statistical Association, Series B*, **64**, 237–252.

Siotani, M., and Wakaki, H. (2006). Contribution to multivariate analysis by Professor Yasunori Fujikoshi. *Journal of Multivariate Analysis* **97**, 1914–1926.

Smith, D. L., Dushoff, J., Snow, R. W., and Hay, S. I. (2005). The entomological inoculation rate and *Lasmodium falciparum* infection in African children. *Nature, letters*, 04024.

Snedecor, G. W., and Cochran, W. G. (1989). *Statistical methods, 8th Ed.* Iowa State University Press, Ames.

Soofi, E. S. (1994). Capturing the intangible concept of information. *Journal of the American Statistical Association* **89**, 1243–1254.

Soule, M. E. (1987). Where do we go from here? *in* M. E. Soule (Ed.) *Viable populations for conservation*. Cambridge University Press, Cambridge, UK. pp. 175–183

Speed, T. P., and Yu, B. (1993). Model selection and prediction: Normal regression. *Annals of the Institute of Statistical Mathematics* **1**, 35–54.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society* **64** (3), 1–34.

Starfield, A. M., and Bleloch, A. L. (1991). *Building models for conservation and wildlife management, 2nd Ed*. Burgess Press, Edina, MN.

Starfield, A. M., Smith, K. A., and Bleloch, A. L. (1990). *How to model it: Problem-solving for the computer age*. McGraw-Hill, New York, NY.

Steidl, R. J. (2007). Model selection, hypothesis testing, and risks of condemning analytical tools. *Journal of Wildlife Management* **70**, 1497–1498.

Stephens, P. A., Buskirk, S. W., Hayward, G. D., and Martinez del Rio, C. (2005). Information theory and hypothesis testing: A call for pluralism. *Journal of Applied Ecology* **42**, 4–12.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* **39**, 44–47.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods*. **A7**, 13–26.

Swihart, R. K., Dunning, J. B., Jr., and Waser, P. M. (2002). Gray matters in ecology: Dynamics of pattern, process, and scientific progress. *Bulletin of the Ecological Society of America* **83**, 149–155.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (Mathematic Sciences) **153**, 12–18 (In Japanese).

Taper, M. L., and Lele, S. R. (2004). *The nature of scientific evidence: statistical, philosophical, and empirical considerations*. University of Chicago Press, Chicago, IL.

Taubes, G. (1995). Epidemiology faces its limits. *Science* **269**, 164–169.

Thomas, L., Laake, J. L., Strindberg, S., Margues, F. F. C., Buckland, S. T., Borchers, D. L., Anderson, D. R., Burnham, K. P., Hedley, S. L., Pollard, J. H., Bishop, J. R. B., and Marques, T. A. (2006). Distance 5.0. Release 2. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK.

Tong, H. (1994). Akaike's approach can yield consistent order determination. Pages 93–103 *in* H. Bozdogan (Ed.) *Engineering and Scientific Applications*. Vol. 1, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Dordrecht, Netherlands.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference* **49**, 137–162.

Umbach, D. M., and Wilcox, A. J. (1996). A technique for measuring epidemiologically useful features of birthweight distributions. *Statistics in Medicine* **15**, 1333–1348.

Van Buskirk, J., and Arioli, M. (2002). Dosage response of an induced defense: How sensitive are tadpoles to predation risk? *Ecology* **83**, 1580–1585.

van der Linde, A. (2004). On the association between a random parameter and an observable. *Test* **13**, 85–111.

Venables, W. N., and Smith, D. M. (2002). *An introduction to R*. Network Theory Limited Publishing, Bristol, UK.

Vieland, V. J., and Hodge, S. E. (1998). Review of *Statistical Evidence: A Likelihood Paradigm*. By R. Royall. *American Journal of Human Genetics* **63**, 283–289.

Vonesh, E. F., and Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. Marcel Dekker, New York, NY.

Wackerly, D. D., and Mendenhall, W., III. (1996). *Mathematical statistics with applications*. Duxbury Press, New York, NY.

Wagenmakers, E-J, Farrell, S., and Ratcliff, R. (2004). Naïve nonparametric bootstrap model weights are biased. *Biometrics* **60**, 281–283.

Wallace, C. S. (2004). *Statistical and inductive inference by minimum message length*. Springer, New York, NY.

Weakliem, D. L. (Ed.) (2004). Model selection. *Sociological Methods and Research*, **33**, 167–304.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Wel, J. (1975). Least squares fitting of an elephant. *Chemtech* Feb. 128–129.

White, G. C., and Burnham, K. P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study* **46**, 120–138.

White, G. C., and Lubow, B. C. (2002). Fitting population models to multiple sources of observed data. *Journal of Wildlife Management* **66**, 300–309.

White, H. (1994). *Estimation, inference and specification analysis*. Cambridge University Press, Cambridge, UK. pp. 380.

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic Press, New York, NY.

Woods, H., Steinour, H. H., and Starke, H. R. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Industrial and Engineering Chemistry* **24**, 1207–1214.

Young, L. J., and Young, J. H. (1998). *Statistical ecology*. Kluwer Academic Publishers, London, UK.

Zellner, A., Keuzenkamp, H. A., and McAleer, M. (2001). *Simplicity, inference and modelling: Keeping it sophisticatedly simple*. Cambridge University Press, Cambridge, UK.

Zhang, P. (1992). Inferences after variable selection in linear regression models. *Biometrika* **79**, 741–746.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology* **44**, 41–61.

Zuur, A. F., Leno, E. N., and Smith, G. M. (2007). *Analyzing ecological data*. Springer, New York, NY.

Index

AICc, 61–64, 66, 68, 72, 84, 86, 93, 96, 113, 119
AIC values, 60, 63, 65, 66, 67, 84, 85
Akaike's information criterion (AIC), 55–58, 60, 61, 62, 64, 65, 68, 70, 146
"All possible models," application, 140
ANOVA model, 8
Approximating model, 56, 57, 59
A priori hypotheses, 6, 9, 16, 78, 96, 116, 118
Asymptotic bias correction, 58
Asymptotic sampling distributions, 64
Averaging detection probability, 115

Bayesian approach, 88, 102, 110, 121, 135, 172, 173
Bayesian information criterion (BIC), 173–176
Bayesian Markov Chain Monte Carlo method, 135
Bayesian methods, 29
Beak lengths, modeling of, 37, 38
Bias-corrected log-likelihood, 57–58
Bias correction term, 57, 61
Bias estimation, in statistical expectations, 170
Bimodality, 37
Binomial coefficient, 169
Binomial distribution, 53
Binomial likelihood function, 36
Binomial model, 41
Boltzmann's entropy, 55, 144
Bootstrap in sampling variance estimation, 129
Bovine tuberculosis, 4, 69

Cement hardening data, 66–67, 112–115, 119, 123
Chi-square distribution, 58
Chi-square test, 43
Cinclus cinclus. See European dippers
Coding theory, 55
Coin flip protocol, 136, 137, 161
Conditional variance, 111, 113
Confirmatory studies, 8–9
Conflict resolution, information-theoretic methods usage, 135, 136
Conflict resolution protocol, 137, 138
Constant hazard functions, 35, 36
Convenience sampling, 20
Correlation coefficient, 34
Count data overdispersion, 126
Covariance matrix, 25
Critical thinking, 10, 12, 15, 17, 37, 47, 49, 73, 77, 78, 158, 164
Cross validation, 72
Cutting edge, in empirical science, 146

Data analysis, 8, 11, 15, 19–24, 59, 62, 63, 98, 112, 118, 144
Data based model selection, 115
Data collection, analysis, 9, 12, 21
Data collection methods, 143
Data dredging, 45, 64
Data fitting, 25
Data information, 11, 20, 31, 60, 106
Data modeling, 21
Data optimization, 36
Data pattern, 46
Data *post hoc* estimation, 146

Data sampling, 14, 98, 116, 120
Data set, 22, 24
Degree of dependence and overdispersion parameter, 127
Degrees of freedom, 7, 43
Design based inference, 8
Deviance information criterion (DIC), 135, 173
Deviance in likelihood-based inference, 167
Discrete random variable, expectation, 169
Discriminant function analysis, 15, 24
Disease transmission, 5
Dominant variable in model selection bias, 129

Effect size, 7
Elephant seals, data overdispersion, 128, 129
Empirical science, 25, 26, 32, 33
Empirical support, 47
Entropy and information theory, 54
Entropy maximization principle, 55
Enzyme inhibition, 44
Enzyme kinetics, modeling of, 44–45
Error distribution, 64, 72
European dippers, 46
Evidence, 2–4, 8, 9, 11, 12, 14–16, 21, 25, 26, 39, 41, 45, 47, 63–65, 78, 83, 85, 87–100, 118, 120, 121, 133, 135, 137–141, 143–146
Evidence ratio, 2, 89, 90, 92, 97, 99, 118, 144, 145
Experimental data, 4
Experimental design, 7, 19, 20

Fast learning strategy, 21
Ferrets, Bovine TB transmission in, 4–6, 23, 35, 69, 93–94, 121
Finch bill lengths, unimodality of, 38
Fisher's likelihood theory, 55, 75
Fitted model, 56
Fixed effects models, 134
Flather's models, 95, 106, 110
Flood effect, on European Dippers, 95–96, 109
Flour beetles, dose response modeling, 41, 42
Force of infection, 4
Fourier series, 25, 121
Freedman's paradox, 16, 132, 145, 165
Full reality, 25, 27–29, 53, 54, 57–60, 62, 66, 70, 75, 84, 85, 88, 99

Gamma distribution, 37, 38, 39, 53, 64
Gender effect, 36
Geospiza fortis, 37
Geostatistical modeling, 71
Goodness-of-fit (GOF), 43, 44, 87, 99
Goodness-of-fit test for data overdisperion, 128, 146

h-likelihood, 135, 159
Hierarchical models. *See* Random effects models
Human medical research, 8

iid assumptions in statistics, 126, 127, 161, 162
Independence, deficiency in count data, 126
Independent and identically distributed *(iid),* 161
Inductive inferences, 6, 20, 21, 98
Inference methods in life sciences, 145, 146
Information loss, 66
Information-theoretic approaches, hen clam experiment, 138–140
Information-theoretic approaches, levels, 146
Information-theoretic approach (s), 11, 47, 55, 63, 64, 72–73, 94–96, 102, 110
Information-theoretic methods, usage in conflict resolution, 135, 136
Information theory, 52, 54, 74

Kullback-Leibler (K-L) information, 53–56, 58, 59, 66, 70, 79, 84, 87, 88, 95, 101, 121, 135, 144, 146

Least squares, 29, 34
Least squares estimates, 68
Least squares methods, 107, 110
Least squares regression, 57, 61
Life sciences, 12, 72, 106
Likelihood-based estimation, 72
Likelihood data function, 41
Likelihood function, estimation, 161, 166
Likelihood methods, problems, 168
Likelihood model, 86–89
Likelihood ratio tests, 58, 166, 167
Likelihood theory, uses, 161, 163, 166
Linearity, 25
Linear/logistic regression in model selection bias, 129
Linear models, 42
Linear regression models, 28

Llikmalc, in hen clam experiment, 138, 139
Logistic regression, 43
Log-likelihood function, 57, 61, 64, 65, 68, 72, 95, 117, 118, 164–166, 168
Log-likelihood of overdispersion parameter, 127, 128

Markov Chain Monte Carlo (MCMC), 172
MARK software, 123
Mathematical model, 21, 25, 72, 88
Mathematical modeling in empirical science, 143
Mathematical statistics, 54, 58
Matrix trace, 70
Maximum likelihood estimate (MLE), 11, 29, 34, 41, 42, 44, 45, 64, 66, 67, 68, 86, 96, 98, 107, 109, 110, 163, 167, 169, 170
Maximum likelihood model, 55, 57
Maximum log-likelihood function, 52, 56, 57, 61, 69, 72, 95, 135, 144
Mean squared errors, 135
Michaelis-Menten model, 44
Missing data, 65
Mixture distribution, 38, 39, 154
Model averaging, 106–109, 112, 113, 115, 131, 132
Model based inference, 9, 15, 21, 25, 26, 88, 120, 176–179
Model *g,* 53, 58, 60, 93, 111
Modeling hypotheses, 35, 36
Model parameters, 26, 56, 86, 99, 109
Model probabilities, 87–93, 95–99, 101, 106–109, 112, 113, 115, 116, 118, 119, 122, 131
Model probabilities *(w_i)* estimation, 134
Model redundancy, 133, 134
Model selection, 9, 10, 59, 62, 95, 97, 106, 109, 110, 115, 116, 118, 119
Model selection bias, 129, 130
Model selection bias and applied data analysis, 130
Model selection bias problems, solutions, 130–132
Model selection bias reduction, model averaging in, 131
Model selection criteria in multivariate analysis, 133
Model selection uncertainty, 62, 63, 70, 84, 95, 109–115
Monte Carlo simulation, 23, 27
MSEs. *See* Mean squared errors
Multimodel inference, 99, 106, 108, 114, 115, 120
Multinomial distribution, 53
Multiple working hypotheses, 3–4, 9
Multivariate AICc, 133
Mycobacterium bovis infection, data of, 23, 24

Nested models, 47, 112
Nested *vs.* nonnested models, 63
Nonlinear regressions, 94
Normal distribution, 53
Null hypothesis testing, 3, 5, 7, 11–13, 46–47, 62, 64, 68, 90, 171, 172
Null model, 47
Numerical methods, 72

Objective bayesian approaches, 11
Observational studies, 10, 11, 15, 35, 44
Occam's razor, 31, 146
Overdispersed count data, 126
Overdispersion *(c)* parameter, in biological data, 127
Overdispersion in count data, coping with, 127, 128
Overdispersion in data on elephant seals, 128, 129
Overdispersion parameter estimation. *See* Variance inflation factor estimation
Overdispersion parameter usages, 128
Overfitting, 25, 30–33, 59, 62, 71, 81

Parameter count K, 71
Parameter estimation, 5, 23, 29–30, 53, 55, 56, 59, 63, 65, 71, 72, 86, 94, 107, 109–112, 114, 118
Parameter heterogeneity in overdispersed count data, 126
Parsimonious model, inference, 24
Parsimony principle, 30, 59, 70, 88, 101
Plasmodium falcipraum infection, 63, 71
Plausible hypotheses, 11
Plausible science hypotheses *(H_i),* 3, 142
Poisson distribution, 53
Portland cement hardening data, 13, 22–23, 91, 112–114, 119
Post hoc, 64, 96–98, 117
Post hoc data analysis, 8, 9, 22, 45
Prediction theory, 108, 109
Predictive likelihood, 56, 68, 72
Predictive mean squared errors (PMSE), 59, 68, 176

Predictor variables, 119, 120
Pretending variable, 65, 117
Principal components analysis, 37
Principle of parsimony, 146
Probabilistic sampling, 7, 20, 23
Probabilistic theory, 19, 23
Probability and likelihood, comparison, 162, 163
Probability distribution, 53
P-values, 5, 6, 13

QAICc usage in overdispersed data, 128
Quasi-experiment, 7, 15
Quasi-likelihood concept, 128

Random effects models, 134, 135
Random sampling, 20–23
Recapture probability *(p),* 47
Regression analysis, 61, 94
Regression models, 107, 120
Relative ranking of importance, 119, 120
Residual sum of squares (RSS), 61
Response curve, dose, 41
Response variable, 108, 109, 119
Reverse jump Markov Chain Monte Carlo approach (RJMCMC), 173
R^2, likelihood version, 168

Sampling protocol, 19, 20, 21
Sampling variance, 110, 111
Schwarz' BIC, 174
Science and technology, 144
Science hypotheses, 11, 12, 25, 26, 62, 63, 64, 67, 68, 71, 72, 84, 85, 87, 90, 92, 93, 94, 98, 99, 107, 121
Second-order bias correction, AICc, 60, 65
Shannon entropy, 78
Size estimation and stratification, 7, 21
Squirrel occupancy, 116
Standard error (se), 7, 30, 67, 113, 114, 117, 118
Statistical expectations, types, 169, 170
Statistical inference, 12, 29
Statistical methods, 29
Statistical science, 6, 8, 10–12, 25, 26, 32, 106
Statistical theory, 11, 54, 98
Stepwise regression, 33
Stochastic biological processes, 62
Stochastic complexity, 25, 62
Stochastic component, cement hardening, 23
Stochastic model, 25, 87

Tail probability, 13
Takeuchi's information criterion (TIC), 57, 70
Tapering effect sizes, 33
Threshold effects, asymptotes, 25
Trace operator, 70
Trade-offs, 30–32, 58, 59, 101
Traditional null hypothesis testing, 46–47
Treatment *vs.* control groups, 7
Tribolium confusu, 41
True distribution *f,* 54
True models, 26, 27

Uncertainty estimation, 87, 90, 95, 109–111, 113
Unconditional variance, 110–112, 114, 115
Underfitting, 30, 32, 59
Unimodal, 37–39, 154

Variance, 7
Variance component, 111, 113
Variance component models. *See* Random effects models
Variance-covariance matrix, 52, 128, 163, 166
Variance inflation factor estimation, 127
Variance, residual, 57, 66

Weibull errors, 25