6

Advanced Topics

Kenneth P. Burnham (1942–) Dr. Burnham received a B.S. degree in biology from Portland State University and his M.S. and Ph.D. degrees in mathematical statistics in 1969 and 1972 from Oregon State University. He has worked at the interface between the life sciences and statistics in Maryland, North Carolina, and Colorado. He has made long strings of fundamental contributions to the quantitative theory of capture–recapture and distance sampling theory and analysis. His contributions to the model selection arena and its practical application have been profound. He was selected as a Fellow by the American Statistical Association in 1990 and promoted to the position of Senior Scientist by the U.S. Geological Survey in 2004. He has a long list of awards and honors for his work, including the Distinguished Achievement Award from the American Statistical Association and the Distinguished Statistical Ecology Award from INTERCOL (International Congress of Ecology). He has just become an elected member in the International Statistical Institute. Ken (left) is shown with Hirotugu Akaike at the 2007 Kyoto Laureate Symposium. Photo courtesy of Paul Doherty and Kate Huyvaert.

6.1 Overdispersed Count Data

Statistical methods are often based on the "iid" assumption: independent and identically distributed data. This assumption is nearly always made in application (time series and spatial models are exceptions); however, in reality, data are often somewhat dependent and not identically distributed. These conditions fall under the concept of *overdispersion*. Count data (zero and the positive integers stemming from some count) are often said to be "overdispersed." There are two issues here. First, overdispersion is a property of the data, not a model; however, overdispersion can be modeled. Second, overdispersion can be modeled as either a lack of independence or parameter heterogeneity. There are a variety of specialized approaches to attempt to deal with overdispersed data, most are at an advanced level and specific to certain problem types. A simple method often serves to lessen the problem with overdispersed count data and it will be introduced in this section.

6.1.1 Lack of Independence

Flipping new pennies and observing the binomial outcomes (i.e., heads or tails) nicely illustrates independence of the outcome from flip to flip. However, count data in the life sciences often have some degree of dependence. Husbands and wives may not be independent with respect to some condition. Individuals in small groups of tadpoles along a mud bank probably die or survive with some degree of dependence within a group. If one tadpole in a group dies it may be that many others die at about the same time and for the same underlining cause. The analysis of count data of litter mates, breeding pairs, schools of fish, and pods of whales should always be suspected of having some degree of dependence. If such count data are analyzed as if they were independent, then the sampling variances tend to be too small (underestimated), giving a false sense of precision (e.g., confidence intervals are too narrow).

6.1.2 Parameter Heterogeneity

Overdispersion can also arise as most statistical methods rely on the concept of parameter homogeneity. Although 200 new pennies may each have essentially the same probability of a head, it is clear that 200 laboratory mice are somewhat variable, almost regardless of the trait of interest. This individual variation leads to what is termed "parameter heterogeneity" and this violates the *iid* assumption. Again, the effect of such heterogeneity, if the data are analyzed under methods that assume parameter homogeneity, is again the underestimation of sampling variances. There may be substantial bias in parameter estimates in some isolated cases.

6.1.3 Estimation of a Variance Inflation Factor

Overdispersion causes the estimated theoretical variances and covariances to be biased low; thus, a first-order approach is to "inflate" these up to a nominal level. This is a simple and often effective procedure. First, we focus on a robust global model; a model with plenty of structure. Here we must assume that there is no structural lack of fit and, therefore, lack of fit can be blamed on overdispersion. This is a strong assumption and one risks the situation where some of the lack of fit is a structural inadequacy of the model and not overdispersion. In this case, the covariances would be inflated (and this might be beneficial) when, in fact, some bias is likely due to inadequate structural modeling.

If there is little reason to suspect some dependence among observations based on counts, then perhaps one should ignore the issue. However, if there is biological reason to suspect overdispersion, then an overdispersion parameter c can be estimated,

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

where χ^2 is the usual goodness-of-fit test statistic based on the global model and df is the degrees of freedom for the test. The overdispersion (c) parameter is also called a *variance inflation factor*. Under the *iid* assumption, $c \equiv 1$. In biological data on counts one often sees $\hat{c}$ in the 1–3 range. Fish in schools, insects in colonies or swarms, or snakes in dens can have overdispersion parameters substantially higher than 4–5. As $\hat{c}$ gets large one must worry that there are structural issues with the model and these are being incorrectly cast as overdispersion.

6.1.4 Coping with Overdispersion in Count Data

Coping with Some Dependence

Used carefully, the estimation of an overdispersion parameter can adjust the analysis in the face of some degree of dependence and parameter heterogeneity. If overdispersion is thought to be an issue and an estimate of the overdispersion parameter is available, e.g., $\hat{c}$, then three things should be done in the analysis (the order is not important):

1. The log-likelihood of the parameters θ, given the data and the model, should be computed as

$$\frac{\log\left(L(\theta \mid x, g_i)\right)}{c},$$

therefore, model selection should use the following modified criterion

$$\text{QAICc}=-\left[2\log\left(L(\hat{\theta})\right)/\hat{c}\right]+2K+\frac{2K(K+1)}{n-K-1}$$

2. The number of parameters (K) is now the number of parameters (the dimension of θ) in the model, plus 1 to account for the estimation of the overdispersion parameter, c
3. The variance–covariance matrix should be multiplied by the estimated overdispersion parameter, $\hat{c}$ (i.e., $\hat{c}$ (cov($\hat{\theta}_i$, $\hat{\theta}_j$) for all the models. Thus, c is used to actually inflate the variance and covariances. Alternatively, standard errors are inflated by the square root of $\hat{c}$.

Once an estimate of the overdispersion parameter has been made from a global model, it is used for all the models in the set (i.e., the three steps outlined above). If $\hat{c} < 1$, then it is rounded up to 1 and no adjustment is made in any of the above quantities. The notation QAICc stems from the concept of quasi-likelihood from a well-known paper by Wedderburn (1974).

The log-likelihood is adjusted in an intuitive way. Usually, the log-likelihood contains all the information in the sample data, given the model and assuming independence. When, instead, there is some dependence, a log-likelihood that assumes independence exaggerates the amount of information in the data. Thus, division by the estimated overdispersion coefficient correctly adjusts the log-likelihood for the degree of dependence reflected in the data.

Highly dependent data have considerably less information and $\hat{c}$ is needed to adjust for the dependence. Assuming everything else is constant, highly dependent data reflect less precision for parameter estimates and selected models with fewer parameters or less structure.

6.1.5 Overdispersion in Data on Elephant Seals

Pistorius et al. (2000) evaluated hypotheses concerning age- and sex-dependent rates of tag loss in southern elephant seals (*Mirounga leonina*) by considering four models. There was belief that these data were overdispersed due primarily to parameter heterogeneity. Burnham and Anderson (2001) made use of these data as an example to explore these issues further. They performed a goodness-of-fit test (*TEST2*, Burnham et al. 1987) on these data, partitioned by gender. The results were

Quantity	Males	Females	Combined
χ^2	157.20	97.92	255.12
df	77	84	161

giving $\hat{c}$ = 255.12/161 = 1.58. This suggests some minor to moderate overdispersion and it is likely to be worthwhile to inflate the variances and covariances and alter the deviance.

Thus, QAICc was used, whereby the deviance was computed as $-2\log(\mathcal{L}(\phi))/\hat{c}$, the parameter count ($K$) was increased by 1 for the estimation of the variance inflation factor, and the covariance matrix for the four models was multiplied by $\hat{c}$ = 1.58. Pistorius et al. (2000) used the bootstrap to obtain estimates of sampling variance and they found the empirical support to be for the models where tag loss was sex- and age-dependent (w_{best} = 0.82) or just age-dependent (W_{second} = 0.18).

As dependence in the data increases, QAICc will tend to select less rich models (i.e., fewer parameters and less structure). This result follows because there is less information when some dependence is present in the data. In a sense, the "effective" sample size is less than *n*. Underdispersion seems hard to imagine; I have not seen this in my experience.

A reviewer brought up the question of independence in time series and spatial modeling problems. Here, the "response variable" is correlated in time or space. Thus, it is the model that attempts to handle the dependencies in time or space (see Renshaw 1991). If successful, the "residuals" will be uncorrelated.

6.2 Model Selection Bias

Technical difficulties can arise when using data to both select a good model and estimate its parameters. Chief among these is the subtle but very important issue of *model selection bias*.

It is difficult for most of us to understand model selection bias because in our regression classes we learned that, given the model, the β_i were unbiased, normal, and have minimum variance. This is all true, *given the model* and its underlying assumptions. In the real world, *the* model is not *given* to us, we must use some analytic approach to select a good model from the data. Again, issues arise when the same data set is used to both select a model and estimate its parameters.

6.2.1 Understanding the Issue

This issue can best be understood in terms of linear or logistic regression. I begin by considering the linear regression function, and for simplicity I will assume all the β_i are positive (and then the discussion relates to overestimation),

$$E(Y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_3).$$

We will define x_1 as a "dominant variable" because its relationship to the response variable (*Y*) is quite important. Here, we might expect β_1/se(β_2) ≈ 3 or 4. Nearly all methods would select this variable to be important (see Miller 2002). The variable x_2 is, in fact, somewhat important, but its relationship to the response variable is a bit weak. Perhaps β_2/se(β_2) ≈ 1. Finally, variable x_3 is tenuous at best. Here, perhaps β_3/se(β_3) ≈ 1/4. The epidemiologist Michael Thun noted, "… you can tell a little thing from a big thing. What's very hard to do is to tell a little thing from nothing at all" (Taubes 1995). This is the concept of tapering effect size (Sect. 2.3.5). Here, β_3 is nonzero and reflects a very weak effect.

Assume we use the linear model above to generate 1,000 data sets where the β_i parameters are known, then we can use a model selection approach (e.g., stepwise, AICc or BIC) to find the best model for each of the data sets. Variable x_1 will likely be selected in virtually all of the 1,000 data sets, while variable x_2 might be selected in perhaps half of the data sets. That is, too little information is contained in many of the data sets and, in view of parsimony, the importance of x_2 and its β_2 is not picked up. More interestingly, it might be that variable x_3 and its parameter β_3 is selected in only a few (i.e., 3–6%) of the data sets.

Now we must ask what are the properties of the estimator $\hat{\beta}_3$ *when it is in the selected model*? Large bias is the answer! The bias arises because about the only time x_3 is in the selected model are cases where it is overestimated. If $\hat{\beta}_3$ is near the actual value (β_3), then the variable does not appear in the selected model. It is only when it happens to be overestimated that it is selected and in the best model. Thus, when one averages across (the few) models where $\hat{\beta}_3$ appears, it is far too large. Thus, a large bias is present and it is this that is called *model selection bias*. The issue extends to variables that are exactly unimportant; i.e., where $\beta = 0$. Occasionally, $\hat{\beta}$ will be large and the inference will be that this variable is important. Model selection bias is not an easy concept but it is both common and important. Model selection bias can often be in the 10–80% range, but can be far more serious (see Miller 2002, for some examples).

When a given data set gives rise to a substantial overestimate, a standard Wald test, $t = \hat{\beta}_3 / \hat{\text{se}}(\hat{\beta}_3)$, would be "highly significant" and x_3 and its parameter estimate of β_3 would be in the model. In this case, the numerator ($\hat{\beta}_3$) is biased high and the denominator ($\hat{\text{se}}(\hat{\beta}_3)$) is biased low, yielding an unreliable test result. AICc and other approaches have this same strong tendency but is harder to demonstrate in an analogous way.

When one has 1–2 dozen predictor variables (i.e., 4,095–16,777,215 models), the opportunity for large biases due to model selection are enormous and the probability of several spurious effects quickly goes to 1. Model selection bias is subtle but its effects are widespread and little understood by many people working in the life sciences.

Model selection bias should be a worry in applied data analysis because the analyst has no way of knowing, from the analysis of a single data set, which parameters might be very much overestimated and which have little bias. In fact, the inference that x_3 is very important is largely spurious. This problem is compounded in that the estimated sampling variances are too low (underestimated), giving a false sense of high precision. Driving this issue is the concept of tapering effect sizes that seem omnipresent in the real world.

6.2.2 A Solution to the Problem of Model Selection Bias

Model averaging offers a solution to the problems of model selection bias (P. Lukacs and K. Burnham, personal communication). This approach applies

TABLE 6.1. Model averaging as a means of reducing model selection bias. The model probabilities are shown at the far right.

1	$\hat{Y} = \hat{\beta}_0$	$+\hat{\beta}_1 X_1$	$+\hat{\beta}_2 X_2$	$+\hat{\beta}_3 X_3$	0.15
2	$\hat{Y} = \hat{\beta}_0$	$+\hat{\beta}_1 X_1$	$+\hat{\beta}_2 X_2$		0.35
3	$\hat{Y} = \hat{\beta}_0$	$+\hat{\beta}_1 X_1$		$+\hat{\beta}_3 X_3$	0.10
4	$\hat{Y} = \hat{\beta}_0$		$+\hat{\beta}_2 X_2$	$+\hat{\beta}_3 X_3$	0.05
5	$\hat{Y} = \hat{\beta}_0$	$+\hat{\beta}_1 X_1$			0.25
6	$\hat{Y} = \hat{\beta}_0$		$+\hat{\beta}_2 X_2$		0.10
7	$\hat{Y} = \hat{\beta}_0$			$+\hat{\beta}_3 X_3$	0.00

to model parameters, not predictions that were covered in Chap. 5. The approach is a type of shrinkage (see below) estimation using model averaging. We will use the case outlined above where x_1 was a dominant variable, x_2 was far less important, and x_3 was barely nonzero. The models with their associated model probabilities are shown in Table 6.1.

There is a "balancing" such that each of the β slope parameters occurs in 4 of the 7 models. Such balancing can be done is one of several ways for many problems, but each parameter should be allowed an equal footing. It is often sufficient to list "all possible models" as a way to achieve the needed balance.

Notice that all the models with the dominant variable tend to have high weights (model probabilities). In contrast, the model with only x_3 has virtually no weight. A robust estimate of each of the 3 β parameters can be made in the usual manner,

$$\hat{\bar{\beta}}_1 = \sum_{i=1}^{7} \mathrm{w}_i \hat{\beta}_{1i} \qquad \hat{\bar{\beta}}_2 = \sum_{i=1}^{7} \mathrm{w}_i \hat{\beta}_{2i} \qquad \text{and} \qquad \hat{\bar{\beta}}_3 = \sum_{i=1}^{7} \mathrm{w}_i \hat{\beta}_{3i},$$

but when a regression parameter does not appear in a model, it is assigned a value of 0. The fact that a parameter does not appear in a model *implies* it has a zero value, and so this ought not seem too surprising upon consideration. Note, in all cases, the model probabilities sum to 1

$$(\text{i.e., } \sum_{i=1}^{7} \mathrm{w}_i = 1)$$

For example, the model averaged estimator for β_3 is $\hat{\bar{\beta}}_3$ computed as a simple weighted average, where zeros (in bold) are assigned for models where the parameter does not appear (see Table 6.2).

That is, $\hat{\bar{\beta}}_3 = \sum_{i=1}^{7} \mathrm{w}_i \hat{\beta}_{3i} = 0.610$. This estimate is below the MLEs as it has been "shrunk."

The fitted equation is a single equation where the parameters have all been model averaged,

$$\hat{Y} = \hat{\bar{\beta}}_0 + \hat{\bar{\beta}}_1 x_1 + \hat{\bar{\beta}}_2 x_2 + \hat{\bar{\beta}}_3 x_3.$$

TABLE 6.2. Computing the model averaged estimate of β_3.

Model	Model weight	MLE of β_3	Product
1	0.15	1.73	0.260
2	0.20	**0.00**	0.000
3	0.25	0.83	0.208
4	0.10	1.42	0.142
5	0.20	**0.00**	0.000
6	0.10	**0.00**	0.000
7	0.00	1.08	0.000
Sum	1.00		0.610

This single equation allows robust predictions to be made, based on each of the regression parameters having been model averaged. This procedure "shrinks" estimates toward zero, and greatly lessens the bias due to model selection. Finally, note that predictions made using this model (above) are identical to those made by making a prediction from each model and then model averaging these. The two approaches are equivalent for linear models.

MC simulations have been carried out to suggest this simple approach is very effective (Lukacs et al. unpubl ms.). Here it is important to use the unconditional variance to account for model selection uncertainty.

Use of "all possible models" is a poor strategy in general; in this specific case using all the models is a simple way to impose a balance and put each variable on an equal footing (e.g., each variable appeared in exactly 4 of the 7 models in this example). There are other ways to maintain this balance; for example, in the data on hardening of Portland cement the single variable models were ruled implausible. Thus, one could get the shrinkage estimates from 11 of the models, rather than using the full set of 15 models and still achieve the needed balance.

Freedman's (1982) paradox is largely resolved using this model averaging approach. The combination of shrinkage model averaging and unconditional variances help guard against spurious results. Freedman (1983) concluded, "To sum up, in a world with a large number of unrelated variables and no clear *a priori* specifications, uncritical use of standard methods will lead to models that appear to have a lot of explanatory power." It seems that this type of model averaging will appear to be useful in lessening these issues. Additional research on this matter will be useful in application. Lukacs et al. (unpublished manuscript) have completed some simulations using logistic regression and found performance to be good in this case. While I do not recommend wholesale data dredging via "all possible models," I believe this type of model averaging will help avoid serious model selection bias when faced with many predictor variables, little science theory, and small sample sizes.

6.3 Multivariate AICc

When one is performing multivariate analyses (e.g., multivariate regression or factor analysis), a slightly altered model selection criteria must be used. Terminology can be confusing; here I am addressing the case where there are more than one response variables (multivariate regression vs. multiple regression – both will typically have several predictor variables). The altered criterion is from Fujikoshi and Satoh (1997) (also see Bedrick and Tsai 1994),

$$\text{AICc} = -\ 2\log(\mathcal{L}) + 2K + 2\frac{K(K+v)}{(np-K-v)},$$

where n is sample size on observations on each of the p variables and v is the number of distinct parameters estimated in the covariance matrix (K includes v). Also see Sclove (1994b), McQuarrie and Tsai (1998:147–149), and Burnham and Anderson (2002:424–426), for additional discussion. In the univariate case, $p = 1$ and $v = 1$ and the criterion above reduces to AICc for univariate cases.

Model selection criteria for multivariate analysis represent an active research area. Recent work includes Siotani and Wakaki (2006) and Seghouane (2005, 2006). It is certainly prudent to work closely with a statistician with expertise in multivariate analysis before going too far with an analysis of complex multivariate data.

6.4 Model Redundancy

Occasionally the model set contains two or more models that are inadvertently the "same." This condition is termed *model redundancy*. Model redundancy does not cause problems with Δ_i values, model likelihoods, or evidence ratios, but the model probabilities (w_i) are affected.

Model redundancy can arise as a mistake or the result of carelessness. Alternatively, a team member might suggest a model where a transition probability (ψ) is modeled as a function of distance (d) from a source as

$$\psi(d) = \frac{\exp\{\beta_0 + \beta_1(d)\}}{1 + \exp\{\beta_0 + \beta_1(d)\}},$$

while another person suggests the model

$$\psi(d) = \frac{1}{1 + \exp\left[-\{\beta_0 + \beta_1(d)\}\right]},$$

These models are actually the same and will have the same $\log(\mathcal{L})$ values. Many models have a chameleon-like form and it is sometimes easy to have 2 or more models with different forms that are actually the same model.

Model redundancy can arise when using semiparametric models (Buckland et al. 1997) where the number of parameters is not fixed, but rather enter from the results of model selection. Consider the parameters in a series of Fourier series being used to smooth and make predictions of weekly storm events discussed in Sect. 5.4. These were nested models where 2 parameters were gained from model to model. Thus, g_1 had 2 parameters, g_2 had 4 parameters, g_3 had 6 parameters, and so on. Model selection indicated that model g_2 with 4 parameters was satisfactory. It is then possible that, without realizing it, the analyst believes he has 4 models with 2, 4, 6, and 8 parameters, respectively. However, the last 3 of the 4 models actually have but 4 parameters (because the additional parameters in models 3 and 4 were not needed (dropped).

Model redundancy can arise in subtle ways and steps should be taken or risk the possibility that the model probabilities will be affected, leading perhaps to inferences that are needlessly poor. The usual effect is that one model gets too much weight. The first and most effective solution is to identify the redundant models and delete them from consideration. This will mean the model probabilities (w_i) must be renormalized, but this is trivial to do.

The alternative solution is to cast the redundant models into a subset and assign the models in this subset an appropriate "weight." For example, let there be 6 models in the set and 2 are found to be redundant (let these be models 5 and 6 for illustration). The solution is to compute the model probabilities using the expression

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)\xi_i}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_r\right)\xi_r},$$

where the ξ_i are 1/5, 1/5; 1/5, 1/5; and 1/10, 1/10. This simple device merely gives the two redundant models 1/2 their usual weight. Note, the $\Sigma\xi_i = 1$ as it must.

6.5 Model Selection in Random Effects Models

This book has been about models that can be termed "fixed effects" models. There is an interesting class of models falling under a general classification of "random effects models;" roughly alternative names include "variance component models," "random coefficient models," and "hierarchical models" (see Vonesh and Chinchilli 1997; Shi and Tsai 2002; McCulloch 2003; and Gurka 2006, for additional details). Because of the design of the data collection, such procedures often allow an inference that is wider in scope than with fixed effects models. These approaches allow separate estimation of a component of variance due to sampling (var($\hat{\theta}$ | model, θ)), distinct from a process variance component (σ^2 = var($\hat{\theta}$)). Here process variation might be temporal or spatial. In a sense, the goal of random effects models is the estimation of population means (μ) and process variances (σ^2) and this can often occur without models with a large number of parameters.

Other useful approaches allow shrinkage estimators; these are estimators that attempt to shrink estimates toward their mean value and the amount of shrinkage depends on the relative magnitude of the sampling variance to the process variance. The number of estimable parameters in shrinkage approaches may not be an integer. Such estimators, as a set, can have smaller mean squared errors (MSEs) than MLEs for the same data. Then there are "mixed" models that allow for both fixed and random effects and these too are being used often. All of these approaches exist at a somewhat advanced level and are seeing increased use in many applied fields in the life sciences. There is a large literature on this important class of models (see Gurka 2006, and references therein).

Model selection for random effects and mixed models can be done under a Kullback–Leibler information framework without modification (assuming the proper likelihood is used and the "number" of parameters is available and correct). Here, standard software provides things such as the RSS or the maximized log-likelihood and AICc can be easily computed by hand. Several software packages now output AIC; however, rarely is AICc provided. The key here is to be sure the correct likelihood is in place.

More complicated hierarchical models (e.g., multilevel hierarchical effects) are very well suited for Bayesian Markov Chain Monte Carlo methods (Gelman et al. 2003; Givens and Hoeting 2005). This is a class of models where Bayesian approaches have a distinct edge over other methods; however, methods based on "h-likelihood" may eventually provide another alternative (Lee et al. 2006). There is a large Bayesian literature on these methods and many applications are beginning to appear in journals in the applied sciences. Model selection in these classes of models seems to rely on DIC, the deviance information criterion (see Spiegelhalter et al. 2002). DIC is implemented in programs BUGS and WINBUGS and is seeing heavy use. DIC has Bayesian roots, comes easily from the MCMC algorithm, and is largely AIC-like in its goals and properties.

6.6 Use in Conflict Resolution

The focus of this book as been on science philosophy and the provision of empirical evidence for *a priori* science hypotheses. This final section mentions the use of information-theoretic methods in the resolution of certain types of conflicts. The key to this application is a "protocol" that is jointly developed as an *a priori* template to guide the resolution of the conflict. This approach assumes that data and data analysis are central to the resolution of the conflict and, therefore, defines a relatively narrow region of potential application.

In all subdisciplines in the life sciences, there are conflicts and controversies and many of these concern technical issues (e.g., does smoking cause lung cancer? does exposure to low levels of lead lower IQ?). Many such controversies exist when alternative economic, social, or legal outcomes hinge on scientific results.

Often as a controversy starts to brew one individual or party will take a partition of the available data, analyze it in a way that suits him, develop the results, and show them to the larger group, expecting them to yield to his position. Unimpressed, others in the group question the choice of the data used, the methods, and the "obviously biased" result. These people then retreat to use the data they feel can be justified, choose their own analysis method, develop what they believe to be *the* results and again expect the larger group to yield to their conclusion. By then personalities are on edge and the controversy may begin to enlarge and become personal. At this stage, the controversy may continue to brew over long time periods or be headed for the courts. It seems better to try to get agreement on the substantive, technical issues; then if the courts get involved it is over the less tangible political issues, but hopefully based on good science.

Here it seems important to clearly separate the science issues (does smoking cause lung cancer or not?) from the management or political implications and the related value judgments (e.g., smoking causes cancer; therefore, ban all smoking products or, at least, tax smoking products heavily). The material to follow suggests a protocol for resolving the science issues in a controversy. Science should ideally provide a uniform result; science results should not align with sponsorship or employment of the scientists. The material in this section is taken largely from Anderson et al. (1999).

Evidence in Conflict Resolution

The overall theme in using information-theoretic methods in the resolution of scientific controversies is to replace the *a priori* set of science hypotheses (H_i) with an *a priori* set of "stakeholder" positions (S_i). In both cases, similar issues are important: careful definition of the problem, good data, sound analysis methods, quantified evidence, and synthesis, usually followed by value judgments.

The goal is then to examine the empirical support for each position with an *a priori* understanding as to what is expected under different outcomes.

6.6.1 Analogy with the Flip of a Coin

The protocol is patterned after the flip of a coin to decide a course of action. In a coin flip, there are numerous issues that must be decided and agreed upon prior to the flip, such as (1) who flips the coin? (2) should the coin land on the floor or on the back of one's hand? (3) who gets to choose "heads" or "tails"?

(5) who gets to flip? and (6) most importantly, what is the exact action to be taken if the coin comes up “heads”?

Take the example of Mary and John deciding who will pick up the tab for lunch. Mary and John debate the preliminary issues and mutually agree that John will flip a single coin, that it must land on the floor, and that Mary chooses “heads” while the coin is in the air, and that this outcome means that John must buy lunch (i.e., “heads” = John buys and “tails” = Mary buys). There is clear, deliberate agreement on these *a priori* issues. These issues represent the agreed upon protocol.

The key to the coin flip protocol is that it is clearly unfair or unethical for one party to change their choice *after* the flip! That is, Mary cannot expect to switch to a position “if it is ‘tails’, John must cover the expenses” *after* the coin has landed and the outcome noted. Similarly, she cannot decide not to play, once the coin has landed and she has lost. The parties can argue about the preliminaries, but once these are agreed upon and the coin is “flipped,” they cannot argue the outcome (e.g., “heads,” therefore John must buy). Of course, if agreement cannot be found on, say, who flips the coin, then the matter of lunch expenses must be settled in one of many other ways (i.e., the player decides not to play). A player can withdraw with honor anytime during the development of the protocol. Football and many other sporting events use a coin flip with specific protocols as part of achieving “fair play.”

6.6.2 Conflict Resolution Protocol

The conflict resolution protocol rests on the *a priori* agreement by all parties on

- The questions to be addressed
- The data to be analyzed,
- The specific data analysis methodologies
- Who performs the analysis
- What outcomes provide evidence for which stakeholder position (to favor one side or the other or remain ambivalent)
- How these outcomes will be announced, reported, or reviewed

The fundamental idea is to argue points and eventually agree on the relevant data (which cannot be changed after results are known), an analysis protocol, and then agree on the interpretation of the results within certain limits. This final point (interpretation) attempts to avoid any ambiguity where both sides argue, after the analysis has been completed, “that proves what I said.”

Management implications (the nonscience) based on empirical results (the science) may often be open to discussion and intense debate. The protocol advocated deals only with the science of the matter. The protocol outlined provides a useful, general framework to deal with the synthesis and analysis of empirical data where decisions are to be based on empirical data (“the best available science”).

Often, synthesis of empirical data for management decisions comes from disparate sources with differing analytical methodologies and interpretations.

Such reviews often list tables of results from different published and unpublished studies from which conclusions are made. However, this approach is often hampered by the different analytical methods used in the separate studies. The approach suggested here differs from those approaches in that there is a deliberate attempt to unify separate studies under a single analytical philosophy and framework that was agreeable, *before the results were in*, to all parties involved. This procedure allows the synthesis of empirical data to have greater scientific credibility and clearly demands consensus among the parties involved regarding methods of data collection and analysis.

Using this protocol might often avoid acrimonious and expensive judicial hearings to arbitrate controversies. In such hearings, both parties present evidence to support only their own, often vested, position. These hearings often aggravate controversy, widen disagreements, and confuse the evidence. While the judicial model has several advantages, I suggest that scientists and managers should attempt an objective resolution of scientific issues, rather than turn over these technical tasks to opposing teams of attorneys and a judge.

Often, relevant data, proper analysis methods, and the interpretation of results in terms of management are disputed by the parties. Issue resolution requires numerous features in our protocol, including involving outside parties with minimal vested interest in the outcome, and *a priori* consensus on directions to proceed. Such issues would benefit from the application of the protocol, once those directions are established. I stress that there is a good deal of flexibility in the application to other situations as long as the philosophical core remains intact.

6.6.3 A Hypothetical Example: Hen Clam Experiments

Anderson et al. (2001) illustrated the use of information-theoretic approaches using a hypothetical experiment to examine the effects of a chemical on monthly survival probabilities of the hen clam (*Spisula solidissima*). A registered chemical (*Llikmalc*) was applied aerially across aquatic habitats for mosquito control and controversy arose over its unintended effects on other aquatic organisms. The hen clam became the subject of conflict between (a) the manufacturer and distributor of *Llikmalc*, (2) the state regulatory agency, and (c) an environmental group.

The protocol was followed in this hypothetical example and the three stakeholder positions (S_j) were obvious from the beginning:

S_1: There is a trivial difference in monthly survival probability and this variation cannot be attributed to the application of *Llikmalc*. There is no treatment effect.

S_2: There is a substantial acute survival effect due to the treatment, lasting one month following the aerial application of *Llikmalc*.

S_3: There is a substantial acute survival effect due to the treatment, lasting one month, followed by a month-long chronic survival effect.

TABLE 6.3. Summary of the evidence for the controversy over the chemical *Llikmalc* and its effect on hen clams.

Stakeholder position	Log $\mathcal{L}$	K	Δ_i	Model probability
S_1	−6,140.52	10	19.63	0.000
S_2	−6,108.51	11	5.40	0.063
S_3	−6,096.26	12	0.00	0.937

The agreed upon protocol made it clear that if a stakeholder's position was unsupported, then the others expected that party to yield. One can see that the stakeholder positions are analogous to the science hypotheses. Here, models must be developed to represent each stakeholder position. Here, an important aspect is that each stakeholder is free to derive their own model to best represent their position. In fact, they might be encouraged to hire expertise in this area. This approach is far different that trying to get all the stakeholders to agree on a single model.

There were many complications in the hen clam study (e.g., overdispersed data, at least 2 different approaches to modeling the recapture probabilities, replicates). I will show only enough of the results to illustrate the type of results one might expect (Table 6.3).

Here it is clear from Table 6.3 that stakeholder position 1 is essentially without support (model probability is 0.000055) and its proponent is expected to yield his position. Most of the support is for stakeholder position 3; the evidence ratio $E_{2,3}$ = 14.8 might be viewed as moderate support for the chronic effect. One could examine the estimate of the chronic survival effect and its precision and make further judgments about the importance of chronic effects. The importance of acute effects have been clearly established with moderate evidence concerning a further chronic survival effect. Model averaging could be done to best estimate both acute and chronic survival effects.

The analysis under an information-theoretic approach is the easy part in conflict resolution; it is getting opposing parties to agree on a fair protocol that is often the challenging part. Still, the underlying driving force is the fairness implied in a coin flip, assuming parties that there will be no surprises, and getting them to understand that this is their opportunity to demonstrate to the others that their position is clearly justified and, therefore, will have strong empirical support. There must be a clear statement as to what is expected in the event of various outcomes. I have been part of a large team of people that have used variations of this approach on the northern spotted owl – old growth forest controversy (see Anthony et al. 2006). This has been North America's largest environmental controversy, spanning some 3–4 decades and the data set has been valued at approximately $40M. The protocol has worked well and it is being continually refined. Details of the protocol development and (multimodel) results are given in Anthony et al. (2006).

This approach can be expected to be useful in only a small proportion of existing conflicts. If people are hired to pressure for a particular position against all reason or evidence, this approach will clearly not work and the issue might as well go on to the courts for partial resolution in the long term. Further, if data are not central to the issue, then this approach will not work. However, I think there are many conflicts or controversies within groups of scientists or managers where this approach has potential. Freddy et al. (2004) present a case where the approach might be judged to be useful, but there were difficulties and compromises.

I have seen controversies where some stakeholders withhold their judgment on an ongoing study until the results are in (the coin lands and is inspected). Then, depending on the result, they are either supportive and in agreement or wildly opposed to everything done by the study group. This scenario should be carefully avoided. The use of information-theoretic approaches to aid in the resolution of conflicts is just one application outside the science realm. While this is primarily a primer on science applications, there are a host of other applications in other arenas that are important.

6.7 Remarks

Heuristically, $\hat{c}$ adjusts sample size downward in the face of overdispersion (K. Burnham, personal communication). For count data the log($\mathcal{L}$) can be written as

$$\sum n_i \log(\pi_i) = n \sum \left(\frac{n_i}{n} \log(\pi_i) \right),$$

where $n = \Sigma n_i$ = sample size. Thus, effective sample size is taken as $n_c = n/\hat{c}$, where $\hat{c} > 1$ and $n_c < n$.

The importance of model selection bias is hard to fathom for a person new to this issue. In some areas of science, I think nearly half of the research work involves the "three demons" – many predictor variables, little science to guide the data collection and modeling, and small sample size. If the results of such work were merely worthless it might not be so bad. However, the results are actually deceiving in that bias suggests the importance of things that are actually not important (i.e., spurious). Ideally, there needs to be a greater awareness of model selection bias and its importance.

There are two cases where "all possible models" finds useful application. First is the ranking of relative variable importance. Second is in computing shrinkage estimates to lessen model selection bias. In both applications, there is a need to put variables or parameters being summed or averaged on an equal footing. In these cases, inferences are not being drawn from the careless, unthinking consideration of "all possible models." Instead, "all possible models" is a device to achieve a proper balance as an intermediate step in a particular analysis type. Beyond these two exceptions, one should not run all the possible models as this is poor practice.