2
Data and Models

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

2.1 Data

Data should be taken from an appropriate probabilistic sampling protocol or from a valid experimental design, which also involves a probabilistic component. These are important steps leading to a degree of scientific rigor. Such

data often arise from probabilistic sampling of some kind and are said to be "representative." Outside of this desirable framework lie populations where such ideal sampling is largely unfeasible. For example, human populations are often composed of members that are heterogeneous to sampling. Thus, by definition, it is impossible to draw a random sample and such heterogeneity can lead to negative biases in estimators of population size. Estimators that are robust to such heterogeneity have been developed and these approaches have proven to be useful, but the standard error is often large. In general, care must be exercised to either achieve reasonably representative samples or derive models and estimators that can provide useful inferences from (the sometimes unavoidable) nonrandom sampling.

Unfortunately, it has been common in some subdisciplines to take data via what has been called "convenience sampling," that is, data are taken from roads or sidewalks or in other "convenient" ways (e.g., near a parking lot or under the shade of a tree). I believe these approaches violate accepted science practice; certainly there is not a valid basis for an inductive inference. All that might be validly said is something about only the sample itself. For example, a conclusion might be "I counted the number of birds I saw along 12 roads in western Ohio and 10% were raptors. Here, nothing can be said about birds in western Ohio in general or about the percent of the birds that were raptors as an inductive inference to some well-defined population. This situation is little different then a child who reports, "I saw some squirrels" – this sort of activity never seems to lead to a new theory or an important discovery. We know a great deal about proper data collection. There are dozens of books on sampling protocols and experimental designs. There is little excuse for getting this issue seriously wrong.

Another common error is the use of the so-called index values as the response variable. Under this approach, the response variable of interest is not recorded, rather it is replaced by a crude index value. Such index values are usually a raw count or some sort of averaging of such counts. These numbers are recorded and "analyzed." Much has been written about the use of index values and I think the evidence is conclusive that they represent an amateur, unthinking approach and is not scientific. The word "data" has the connotation that there is recoverable information in the data; index values are not data, they are just numbers. None of the procedures in this primer claim to make sense out of nonsense. If the data have not been taken with care, using proper procedures, then the so-called findings will likely be only an assortment of uncertainty and disinformation. DeLury (1947), a famous fisheries biologist, asked, "Is an untrustworthy estimate better than none?" Meaningful data of sufficient quantity are the grist of scientific bread.

There are two conceptual aspects. First, is the study sound so that an inductive inference can be justified? Second, are the data analysis methods sound? The first is not a data analysis issue, rather this question asks if the science of the matter is reasonably well in place and if the data have been collected in a reasonable manner. The second relies on adequate modeling and on objective

approaches to model selection (Chap. 3). We must try to guard against rushing too quickly to data analysis, when the subject matter science is still underdeveloped or if the data are seriously compromised. The science question should be carefully thought out and plausible hypotheses derived. These matters represent hard work and must typically take thought over a period of many weeks or months. The success, in the end, will rest on these science issues being well done – we must not think the analysis will somehow make up for serious inadequacies during these initial steps. These issues will never live up to the ideal; thus, the concept of evolving sets of hypotheses often prove very useful and lead to an effective strategy for fast learning.

In serious work, data are carefully collected during a pilot study. The pilot study allows investigators the chance to work out the bugs in field or laboratory application and attempt some degree of optimization of the sampling protocol or experimental design to be used. Required sample sizes are estimated, stratification is considered, etc. Engineers routinely conduct feasibility studies before they begin a project and life scientists should take a similar approach before the actual data collection begins. If resources are found to be inadequate for the task, it is often better to wait until the needed resources are assembled before beginning the project. Such waiting allows time for additional planning and refinement while gathering the resources needed.

If data are collected in an appropriate manner, then there is *information* in the sample data about the process or system under study. In simple cases with continuous data, some of this information can be retrieved and understood using graphs (e.g., X vs. Y), plots, histograms, or elementary descriptive statistics (e.g., estimated means and standard deviations). However, in nearly all interesting cases, a mathematical model is required to retrieve the information in the data and allow some understanding of the system. Scientists often want to make a formal inductive inference; that is, the process of going from the sample data to an inference about the population from which the sample was drawn. Deductive statements can usually be classed as either valid or invalid. However, inductive statements (inferences) are not made with certainty and inferential statements can range from very weak to very strong. Inductive inference concerns weighing evidence and judging likelihood, not proof itself. Statistical science has allowed the inductive process more rigor and the ability to address a deeper level of complexity. Nearly all questions in life sciences seem to be inductive.

Valid inductive inference is another example of rigor in science but it rests on certain important requirements. This leads to "model based inference" and this has had a long history in the sciences. The *inference* comes from a model that approximates the system or process of interest. In some sense we can think of a properly selected model as the inference (at least for inferences made from a single model).

Investigators should continue to think and rethink the collection of working hypotheses as the data are collected. This is a place to delete a hypothesis in the set because of field evidence or because it seems unfeasible to measure variables that are central to a particular hypothesis. This is a time to add new

hypotheses or refine existing hypotheses. At any given time, our knowledge is based on hypotheses that have shown their competitive fitness by surviving to this point; a "survival of the fittest" as hypotheses struggle for continued existence. There is a competitive struggle that eliminates hypotheses that are unfit (found to be implausible, based on one or more data sets). For example, perhaps elevation is not important to a response variable, as first thought, rather it is actually temperature (which is negatively correlated with elevation). Should some interaction terms be considered due to observations in the laboratory? Does it seem that two variables might be very negatively correlated, suggesting care will be needed to understand this? The focus should remain on the candidate set of science hypotheses; ideally, these should be fixed once the analysis begins. Following these activities, some tentative hypotheses might be added *post hoc*, but such results must be treated more carefully.

Some common sense and art are involved in the concept of an evolving set of hypotheses. Sometimes it might be premature to delete some hypotheses if the data set is small; in such cases perhaps judgment should be reserved until another data set is available. In analyzing data from small samples, one must guard against dismissing some larger models with more structure because a new and larger data set might be able to support the additional structure. Such judgments can be guided by methods given in Chap. 4.

2.1.1 Hardening of Portland Cement Data

Our first example will be the data on four explanatory variables thought to be related to cement hardening. The meager ($n = 13$) data are shown in Table 2.1.

TABLE 2.1. Cement hardening data from Woods et al. (1932). Four variables (in percent), x_1 = calcium aluminate ($3CaO{\cdot}Al_2O_3$), x_2 = tricalcium silicate ($3CaO{\cdot}SiO_2$), x_3 = tetracalcium alumino ferrite ($4CaO{\cdot}Al_2O_3{\cdot}Fe_2O_3$), x_4 = dicalcium silicate ($2CaO{\cdot}SiO_2$), are given with the response variable, y = calories of heat evolved per gram of cement after 180 days of hardening.

x_1	x_2	x_3	x_4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

In this case, the sample size is 13 and this must be considered to be generally inadequate. One is taught in STAT101 that a sample of 25–30 is often required just to estimate a simple mean from a sample of a population that is approximately normally distributed. Kutner et al. (2004) recommend 7–10 observations for each predictor variable. Thus, we have to be realistic when our interest lies in the estimation of more complicated parameters such as finite rates of population change (λ_i), or an enzyme inhibition rate (f), or some hazard rate ($h_{(t)}$). The cement hardening process is likely to be somewhat deterministic with a fairly weak stochastic component. Thus, even with the small sample available, perhaps some interesting insights can be found in this example. Note that both the response variable and the predictor variables are continuous; the response variable is unbounded while the predictor variables are percentages and bounded between 0 and 100.

The basis for a valid inductive inference in this example rests on the various chemical compounds being reasonably uniform. That is, dicalcium silicate ($2CaO{\cdot}SiO_2$) is the "same" from place to place. Thus, random samples would produce little variation in this variable or, for that matter, the other three variables. This is an important step or generalized inference from these data will be compromised.

Large sample size conveys many important, but sometimes subtle, advantages in the statistical sciences. Large sample size carries more information and such information is a major focus of this primer. Investigators should make every attempt to garner the resources to allow an adequate sample size to be realized. There is a large literature on the establishment of sample size, given either some background data from a small pilot survey or outright considered guesses about the system to be studied (see Eng 2004). Monte Carlo simulation studies provide another means to predict the sample size for a particular application (see Muthen and Muthen 2002).

2.1.2 Bovine TB Transmission in Ferrets

The second example is the data on disease transmission in ferrets in New Zealand (Table 2.2). While the sample size here is moderate (n = 319), estimates of the per year force of infection ($\hat{\lambda}$) varies by a factor of 88; thus it seems realistic that the data might be adequate to reveal some interesting insights. High variances seem to be the rule in many areas of life sciences, making data analysis challenging and making inferences somewhat tentative in many cases because of the uncertainty. Increased sample size can often help combat these issues.

These data consist of counts and are therefore of a substantially different type than the data in the earlier example. Ferrets were caught in baited traps systematically placed in selected areas (based on prior information from wildlife surveys or from tuberculin testing of cattle herds). Traps were checked daily over a 5–10-day sampling period. We might ask if the data came from a strict probabilistic sampling frame – no, probably not. Animals willingly

TABLE 2.2. The data on infection and the estimated force of infection ($\hat{\lambda}$) of *Mycobacterium bovis* infection using modified exponential models (from Caley and Hone (2002).

Site	Gender	No. examined	No. infected	$\hat{\lambda}$/ year
Lake Ohau	M	57	3	0.19
	F	54	2	0.09
	Total	111	5	0.14
Scargill Valley	M	37	5	1.40
	F	39	8	0.65
	Total	76	13	1.02
Cape Palliser	M	15	11	2.69
	F	23	10	1.24
	Total	38	21	1.97
Castlepoint	M	27	21	7.90
	F	21	10	3.65
	Total	48	31	5.77
Awatere Valley	M	24	16	4.64
	F	22	12	2.15
	Total	46	28	3.40

or unwillingly get trapped and there is surely heterogeneity in individual trapability. Are the sample data "representative" to allow a valid inductive inference to the population of interest? I suspect so; however, the authors should ideally make this argument.

2.1.3 What Constitutes a "Data Set"?

There is sometimes confusion as to what represents a "data set" in the literature (e.g., Stephens et al. 2005). There are few restrictions on a data set as long as its components have information on the same issue of interest.

Questions concerning the extent of a data set can often be answered by examination of the response variable. In a treatment-control experimental setting, the response variable might be a concentration level of a compound in a blood sample; thus, one data set because both data sets have information on the same issue of interest. However, if different response variables are used (different issues of interest) across some categories, then these constitute different data sets.

I will provide some examples that might help people understand this matter. Consider a simple treatment and control study; one might think there are two "data sets" here, one for the control and another for treatment. Not so – this is to be treated as a single data set. Consider a discriminant function analysis with seven discriminator variables with the analysis being to find out which subset of the seven might serve in a parsimonious model for inference about the discrimination. Here the "data set" consists of the binary response variable and the seven discriminator variables. Of course, given several models in the set, only one (at most, assuming a global model) will have all seven variables

in it. Other models will have fewer than seven variables, but this does not invalidate the notion of the "data set" (see Lukacs et al. 2007).

2.2 Models

Quantification is nearly essential in the empirical sciences where stochasticity is substantial, there are several different sources of variability (factors), or there is some degree of complexity. This complexity might arise from multiple variables, interactions between and among variables, high variability, nonlinearities (e.g., threshold effects, asymptotes), and a host of other issues. Unless one is engaged in simple descriptive studies, they must deal with mathematical models. Such is certainly the case I focus on here – model based inference. We should not think of this requirement as negative; instead, quantification allows both rigor and the ability to better understand far deeper science issues. Soule (1987:179) offered, "Models are tools for thinkers, not crutches for the thoughtless." Box (1978:436) records that R. A. Fisher felt some statisticians were trained strictly mathematically and that many of them seem to have no experience of the valuable process known as "stopping to think."

We are not trying to model the data; instead, we are trying to model the information in the data. The goal is to recover the information that applies more generally to the process, not just to the particular data set. If we were merely trying to model the data well, we could fit high order Fourier series terms or polynomial terms until the fit is perfect. Data contain both information and noise; fitting the data perfectly would include modeling the noise and this is counter to our science objective. Overfitting is a poor strategy and it goes against the notion of parsimony, a subject to be addressed shortly. Models are *central* to science as they allow a rigorous treatment and integration of:

- Science hypotheses (the all important set $\{H_i\}$)
- Data (e.g., continuous or discrete or categorical),
- Statistical assumptions (e.g., Weibull errors, linearity)
- Estimates of unknown model parameters (θ) and their covariance matrix Σ

Models are only approximations to full reality. Box (1979) said "… all models are wrong, some are useful." We should think of the value of alternative models as better or worse, instead of right or wrong. While a driver's license is "valid" or not, models do not share this property. The strength of evidence for competing models is very much central to both science and this textbook.

Models must be derived to carefully represent each of the science hypotheses. These models are always to be probability distributions. The idea is that each hypothesis has a model that fully represents it; then we can think of hypothesis i and model i as almost synonyms. That is, the goal is to have a one-to-one mapping between the ith hypothesis and the ith model:

$$\boldsymbol{H_1 \Leftrightarrow g_1, H_2 \Leftrightarrow g_2, \ldots, H_R \Leftrightarrow g_R}.$$

People in the life sciences are often poorly trained in modeling techniques; this might be a place where the investigator will want to seek advice or collaboration with a person in the statistical sciences.

Then the science question asks, "What is the support or empirical evidence for the *i*th hypothesis (via its corresponding model), *relative to others in the set*. This leads us to the "model selection" problem. So, finally the issue becomes the *evidence* for each of the hypotheses (and their associated models), *given the data*. Of course, hypotheses and their corresponding models not in the set are out of consideration until, perhaps, they are added at a latter time as the set evolves. So, now we can ask if hypothesis C is 10 times as *likely* as hypothesis A? Is the support for hypotheses A and B nearly equal? Is hypothesis A 655 times more *likely* than hypothesis D? If so, would we take this as very strong evidence? These are the types of science issues that can be answered easily using the existing theory for model based inference.

Often inferences are based on the best hypothesis in the set. While "best" is not defined until Chap. 3, standard analysis often tries to determine or estimate which of the hypothesis is the best, based on the data. Inference is then based on this hypothesis, via its corresponding model – *model based inference*. Assuming models have been derived to represent the hypotheses in the set, this is the so-called model selection problem. A "good" model is able to properly separate information in the data from "noise" or noninformation. Finding such a model is a generic goal of model selection. Now I begin to use the concept that a science hypothesis and its model are (ideally) synonymous.

Many standard approaches to model selection have been developed, including adjusted R^2; Mallows' C_p; step-up, step-back, and stepwise regression, to name a few. As one might expect, the early approaches are rarely the best ones; what is not expected is that the early methods are still being taught in mainstream statistics classes (at least for nonstatistics majors) and readily available in the most well-known statistical computing packages. Most selection approaches (e.g., stepwise regression) are based on some sort of theory but they are often not based on any underlying theory concerning what is a good *fitted* model, given the data; hence, no rigorous criterion of "best" model. The methods do not have a proper underlying theory, just a semblance of semirelevant theory. The model (or hypothesis) selection issue is central to data analysis: "Which hypothesis/model should I use for the analysis of a particular data set?" and "How can this be best done?"

Approaches are needed to provide quantitative *evidence* for the hypotheses in the set. As information can be quantified in various ways, approaches have been recently developed to address the model selection problem as well as an empirical ranking of the hypotheses in the set, through the associated models. Here it starts to become clear that the modeling step is nearly as important as the hypothesizing step in empirical science.

2.2.1 True Models (An Oxymoron)

Models are never "true"; models do not reflect reality in its entirety. In the real world with real data, there is no valid concept of a model that is exactly true, representing full reality. Models are approximations by definition if nothing else. If we had a true model, we would still have to estimate its many parameters and try to interpret the complex result. Any such true model would be quite complicated and involve a great many parameters. Thus, an extraordinarily large sample size would be required, unless it is also assumed that the true model somehow came with its true parameters known to the investigator! I find it hard to imagine a situation where the researcher knew the exact functional form of the true model *and* all of its parameter values! Some scientists might take the view that any such "true" model must be considered infinite dimensioned; perhaps, this is a useful concept but it is just another way of saying there is no valid notion of a true model. Recently I have seen the term "inexact models" used; I believe all models are approximations and, therefore, "inexact."

Computer simulation studies often use Monte Carlo methods to simulate "pseudodata" from a mathematical model, with parameters known or given. Here the exact form of the model and its parameters are known – this is properly termed a generating model. In this computer sense, the "true model" and its parameters are known. A common mistake in the statistical literature has been to provide many replicate pseudodata sets from a generating model, include this model in the set, and then proceed to ask questions about which model selection method most often selects the generating model. Such circular results are of little use in the real world where data arise from complex (and only partially observable) reality, not from a simple parametric model. Real data do not come from models and selection criteria that are designed to select a so-called true model are misguided.

Going further, there is the notion of "consistency" in model selection. Here, some procedures are classed as being "consistent," meaning that as sample size increases (often by three to five orders of magnitude) the probability of selecting the "true" model approaches 1, given the true model is in the set (see Appendix D). This concept seems strained if either the true model is not in the set or if the "true" model is infinite dimensional. In reality, the model set changes as sample size is increased by orders of magnitude and this makes the notion of consistency strained.

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

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

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

Approximating models share some features with maps. Maps fail to capture every detail on the landscape, regardless of their scale. Both data and maps contain errors of omission; this seems unavoidable. Errors of commission should, in principle, be avoided. A map should not show a road or stream that does not exist, while we should not find an effect in the data that does not exist (a spurious effect). All maps are wrong, but some are useful, at least in certain contexts. A map of Switzerland is of limited use in the United States, but might be very useful in Switzerland. Of course, there is no true map.

2.2.2 The Concept of Model Parameters

In many cases, parameters are real entities. For example, the size of a population of parrots in an aviary can be determined by a census at a given point in time and this count is a parameter (N, the population size). If we have a time series of censuses of this population, the parameters are N_t, where t is time. However, parameters are often human constructs and are important in understanding systems or processes. The probability of death in a fish population in a large lake is unobservable and not really a parameter in some sense. Instead, we, as investigators, define an arbitrary time interval (such as a month or a year) and derive models that include the probability of death as a parameter, and proceed to estimate this. It is a model parameter, but is it a parameter intimately associated with the population of fish? Perhaps not? Linear regression models are merely first-order approximations to often complex processes of interest. Any particular β (regression "slope") is unlikely to be a parameter associated with the process itself. Similarly, λ, the finite rate of population change, is hardly a parameter that can be directly observed or measured, but it serves as a very useful construct in population ecology. Scientific understanding can often be aided by the notion of parameters, whether real or just useful or directly observed or unobservable.

In fact, thinking that truth is parameterized is itself a type of (artificial) model based conceptualization. Going deeper, mathematics itself is a "model" when used to represent reality or concepts or hypotheses. Mathematics is a human construct and does not exist in the same sense as reality. Sometimes it is useful to think of f as full reality and let it have (conceptually) an infinite number of parameters. This "crutch" of infinite dimensionality at least keeps the concept of reality even though it is in some unattainable perspective. Thus, $f(x)$ represents full truth, and might be conceptually based on a very large number of parameters (of a type we have not even properly conceived) that gives rise to a set of data x.

Akaike noted that the success of the analysis of real data depends essentially on the choice of the basic model. Successful use of statistical methods depends on the integration of subject-matter science into the statistical formulation. This demands a significant amount of effort for each new problem. This is where the science of the issue enters consideration: a major step.

2.2.3 *Parameter Estimation*

It is a *fitted* model that is the basis for statistical inference; hence, parameter estimation is very important. If the sample size is small, the parameter estimates will typically have large variances and wide confidence intervals and might be so uncertain as to be of little use. Large sample size conveys many important advantages in terms of parameter estimates and model selection.

Given a model and relevant data, procedures were developed nearly a century ago to estimate model parameters. Three common approaches have emerged for general parameter estimation: least squares, LS (or "regression"), maximum likelihood, ML, and Bayesian methods. Least squares has been popular; however, its domain is primarily the class of the so-called general linear models (e.g., regression and ANOVA). I will say little about this approach. The much more general approach is Fisher's maximum likelihood (see Appendix A). The notion of ML is compelling – given a model and data, taking as the estimate the value of the parameter that is "most likely." Hence the name maximum likelihood estimate (MLE); it is the value of the parameter that is most likely, given the data and model.

As sample size increases (asymptotically), MLEs enjoy several properties (within certain regularity conditions): unbiased, minimum variance, and normally distributed. In addition, if one takes an MLE and transforms it to another estimate, it too is an MLE (the "invariance" property). These are important properties and explain partially why likelihood is so central to statistical thinking.

An important component of data analysis relates to the fit of a model to the data. These activities focus primarily on a global model and include such things as a formal goodness-of-fit test, adjusted R^2 value, residual analyses, and checking for overdispersion in count data. If the global model is judged to be "poor," then further data analysis will likely be compromised.

A person new to statistical thinking often finds it difficult to relate data, model, and model parameters that must be estimated. These are hard concepts to understand and the concepts are wound into the issue of parsimony. Let the data be fixed and then realize the information in the data is also fixed, then some of this information is "expended" each time a parameter is estimated. Thus, the data will only "support" a certain number of estimates, as this limit is exceeded parameter estimates become either very uncertain (e.g., large standard errors) or reach the point where they are not estimable.

2.2.4 Principle of Parsimony

A model has structural and residual components. Parsimony relates to under- and overfitting models. Examination of the graph in Fig. 2.1 shows that an underfitted model (the left side of Fig. 2.1) risks not only high bias, but also the illusion of high precision ("a highly precise wrong answer"). Underfitting relates to the case where some model structure is erroneously included in the residuals. Of course the investigator does not necessarily know the situation she is in. Overfitted models are also to be avoided because further examination of the graph suggests that overfitting (the right side of Fig. 2.1) risks including too many parameters (that need to be estimated) and a high level of uncertainty. Overfitting relates to the case where some residual variation is included as if it were structural. This may seem like the lesser of two evils;

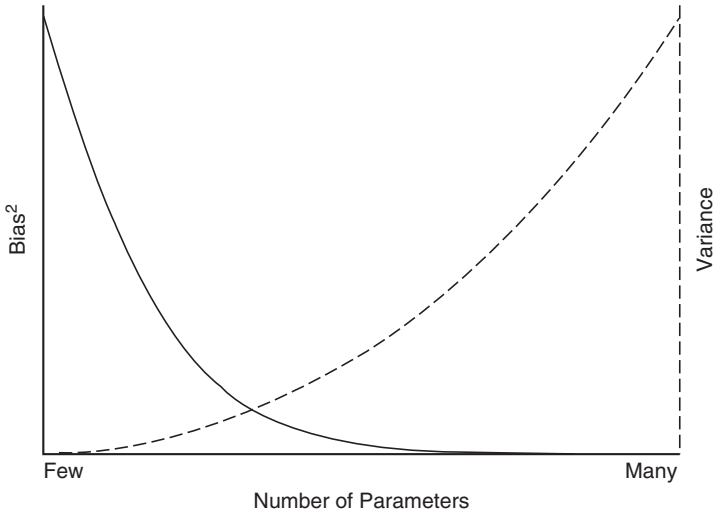

FIG. 2.1. The Principle of Parsimony is illustrated here as a function of the number of estimable parameters (K) in a model. There are two processes here: first, bias (or squared bias) declines as K increases and, second, the variance (uncertainty) increases as K increases. These concepts suggest a trade-off whereby the effects of underfitting and overfitting are well balanced.

however, precision is lessened, often substantially. Overfitting implies that some noise (noninformation) has been included in the structural part of the model and the effects are not part of the actual process under study (i.e., spurious). Edwards (2001:129) says it in an interesting way,

> *"…too few parameters and the model will be so unrealistic as to make prediction unreliable, but too many parameters and the model will be so specific to the particular data set so to make prediction unreliable."*

Clearly, one wants a proper trade-off between squared bias vs. variance or, said another way, between under- and overfitting. Either extreme will result in unreliable prediction. Residuals might be pure noise or information that cannot be decoded yet. The concepts of under- and overfitting depend on sample size; as sample size increases, additional information is available in the data, and smaller effects can be identified. Thus, residual variation can be understood and this transfers to the structural part of the model. Parsimony cannot be judged against any notion of a true model.

The concept of parsimony in modeling and estimation has been an important statistical principle for several decades. The general notion of parsimony has a much longer history in science and engineering and is closely related to Occam's razor. Parsimony is a fundamental issue in science and it is easy to overlook its depth and importance. Occam's statement has a literal translation from Latin, but is commonly referred to as "Occam's razor" meaning roughly to "shave away all that is not needed."

Parsimony appears to be a simple notion; however, it is easy to underrate its importance and its centrality in modeling, model selection, and statistical inference. Parsimony can be viewed as a trade-off between squared bias and variance (variance is a squared quantity, thus bias is squared for some comparability). Think of parsimony as a function of the number of estimable parameters in a model (denote this parameter count K). Given a fixed data set, two things happen as the number of model parameters to be estimated are increased (the standard example is a polynomial where additional parameters are introduced from a linear, to a quadratic, to a cubic, etc.). First, squared bias decreases as more parameters are added – this is good. Second, uncertainty (measured by the variance) increases as more parameters are added – this is not so good (Fig. 2.1).

The addition of more parameters reduces bias but, in doing so, increases the uncertainty. That is, for a given data set and its context, there is a "penalty" or "cost" for adding more parameters that must be estimated. It is the need to *estimate* parameters from the data that is the difficulty. If one could somehow add parameters with *known* values, the situation would be simple: that is, consider only models with a large number of parameters. Unfortunately, parameters in these models are not known; reality is harsh in this regard and parameters must be estimated based on the information in the data. Each time a parameter is estimated, some information is "taken out" of the data, leaving less information available for the estimation of still more parameters.

Parsimony exists near the small region where the lines cross – a trade-off (Fig. 2.1). Parsimony is a conceptual goal because neither bias nor variance is known to the investigator analyzing real data. There are many specific approaches to achieving parsimony but the important concept does not, by itself, lead to a specific criterion or recipe. Parsimony is a property of models (and their parameters that must be estimated) and the data.

There is a large literature admonishing investigators to avoid overfitting as this leads to spurious effects and imprecision. An equally large literature warns of underfitting because of bias and effects that are present, but missed during data analysis. Until somewhat recently, statistical science lacked an effective way to objectively judge the trade-off – how many are too many, how many are not enough. This has been largely resolved for a wide class of problems and is another example of the advantage (actually necessity) of quantification. Rigor in empirical science has a basis in quantification. All methods for model selection are linked in some manner with the principle of parsimony.

I have had biologists state that "A biologically reasonable model is 'punished' because it has too many unknown parameters." Indeed, the estimation of parameters sucks information from the data to the point that little or no information is left for the estimation of still more parameters. It is easy, at first, to think that parameters come somehow "free" and that complex biological models can be developed with little or no data. Instead, the reality of the situation is that parameter uncertainty must harken back to the concept of parsimony. A partial solution to obtaining increased biological reality is to obtain a large sample size or improve study design (e.g., control some factors) as these allow parameter estimates with good precision and functional model forms to be evaluated.

In model selection, we are really asking which is the best model *for a given sample size*. Given a real process that has some realistic degree of complexity and high dimensionality, a high-dimensioned model might be selected as best if the sample was quite large. In the same situation, a small, low-dimensioned model might be expected if the sample was small. A very rough rule of thumb advises that at most $n/10$ parameters can be estimated; thus for observations on a sample of 30 individuals, one might be able to estimate about three parameters (e.g., β_0, β_1, and σ^2) in a regression model. This is often less than what biologists attempt with such small data sets.

Model selection resulting from the analysis of sparse data usually suggest a simple model with few parameters. Such results should not be taken to suggest that the system under study is necessarily simple. On the contrary, if a virtually "null" model is selected, this usually points to an insufficient amount of data to fit anything more realistic. Even then, if the best model is, for example, one with no time effects, one should not infer the process is time invariant. Instead, the correct interpretation is that the variation in some parameter across time is small and such variation could not be identified with the small amount of information in the data.

We are really asking – how much model structure will the data support? A good fit is *not* sufficient, we need predictive ability, and this involves parsimony – how many parameters can be estimated and included in a model? Overfitting risks (by the addition of extra parameters) the inclusion of some of the random "noise" as if it were structure. Model selection criteria allow an objective measure of how many parameters can be fitted to a model, given the sample size. We can chase truth, but we will never catch it and parsimony is central to the chase.

2.2.5 Tapering Effect Sizes

In perhaps all of the empirical sciences, there are a wide range of "effect sizes." There are the large, dominant effects that can often be picked up even with fairly small sample sizes and fairly poor analytical approaches (e.g., stepwise regression). Then there are the moderate-sized effects that are often unveiled with decent sample sizes and more adequate analysis methods. It is more challenging to identify the still smaller effects: second- and third-order interactions and slight nonlinearities. Increasingly large samples are needed to reliably detect these smaller effects. Beyond these small effects lie a huge number of even smaller effects or perhaps important effects that stem from rare events. This situation is common as any field biologist can attest. We say there are "tapering effect sizes" and we can chase these with larger sample sizes, better study design, and better models based on better hypotheses. The notion of tapering effect sizes is everywhere in the real world and it is hard to properly emphasize their importance.

Tapering effect sizes are what preclude the notion of a true model. Just the high-order interactions are quite complex. Consider the ramifications of the various systems in the human body as body temperature climbs to 105° or as one finishes a marathon run. The life sciences are all about a wide variety of tapering effect sizes.

2.3 Case Studies

2.3.1 Models of Hardening of Portland Cement Data

This is a well-known data set and authors typically approach the issue as a multiple linear regression problem with four predictor variables. The global model is

$$E(y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_3) + \beta_4(x_4),$$

where y is the calories of heat evolved per gram of cement after 180 days, x_1 the percent calcium aluminate ($3CaO \cdot Al_2O_3$), x_2 the percent tricalcium silicate ($3CaO \cdot SiO_2$), x_3 the percent tetracalcium alumino ferrite ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$),

x_4 the percent dicalcium silicate ($2CaO{\cdot}SiO_2$), and $E(\cdot)$ is the expectation operator (see Appendix B).

This example problem has two objectives: variable selection and prediction. The analysis could be done in a least squares (LS) or maximum likelihood (ML) framework (Appendix A). The LS and ML estimates of the β_i parameters will be identical; the two estimates of σ^2 will differ slightly. This might be a place for the reader to review quantities such as the residual variance σ^2, residual sum of squares RSS, adjusted R^2, the covariance matrix Σ, various residual analyses, and the notion of a global model.

Because only four variables are available, the temptation is to consider all possible models ($2^4-1 = 15$) involving at least one of the regressor variables. Burnham and Anderson (2002), strictly as an exploratory example, considered the full set of models, including the global model {1234} with $K = 6$ parameters (i.e., β_0, β_1, β_2, β_3, β_4, and σ^2). They generally advise against consideration of all possible models (15 in this example) of the x_i (note that even more models would be needed if interactions, powers of the predictor variables, or other nonlinear relationships were employed).

In contrast, for this example, I will try to limit the set to those that seem plausible, particularly in view of the small sample size. Using all possible models usually represents an unthinking, naive approach. I have already noted that the global model is essentially singular as the numerical values for the four variables sum to approximately 1 (rounding prevents some sums to be exactly 1). Thus, the global model can be dismissed in the example. I already eliminated the four single variable models as cement is a mixture of ingredients. So, now the set is down to 10 models.

Additional thinking (Sect. 1.8) about the chemical similarity of the pair of variables 1 and 3 and the pair 2 and 4 was relevant. Without the curse of data dredging, it is advisable to examine the correlations between these pairs, based on the data available. Such correlation analysis substantiates the observation; the correlation coefficient between x_1 and x_3 was -0.824 and the correlation between x_2 and x_4 was -0.973). Thus, including both variables within a pair would not be advisable, particularly in view of the fact that the sample size is only 13 observations. However, we do not know if x_1 or x_3 is the better predictor, nor do we know if x_2 or x_4 is the better predictor. Thus, the following five hypotheses and variables lead to five models making up the candidate set:

H_1	0 variables	g_1	$E(y) = \beta_0$
H_2	x_1 and x_2	g_2	$E(y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2)$
H_3	x_1 and x_2 and $x_1{*}x_2$	g_3	$E(y) = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_3(x_1{*}x_2)$
H_4	x_3 and x_4	g_4	$E(y) = \beta_0 + \beta_1(x_3) + \beta_2(x_4)$
H_5	x_3 and x_4 and $x_3{*}x_4$	g_5	$E(y) = \beta_0 + \beta_1(x_3) + \beta_2(x_4) + \beta_3(x_3{*}x_4)$

Hypothesis H_1 has no predictor variables and is not in the original 15 possible models. I include it here as an example. Of course, the numerical values for the ML estimates of the β parameters will differ across models (i.e., β_1 "means" different things and is model specific). Now it becomes clear that

hypothesis 4 (H_4), for example, has its corresponding model (g_4). This is a set of first-order models with all the variables entering a linear model. The model set is crude; however, there are little data and so more complex models might not be justified. Note how knowledge of sample size affects the number of parameters that might reasonably be estimated; this requires some experience. However, even a C student just finishing a class in applied regression would surely not attempt to estimate 8–10 parameters from this data set. This will serve as our initial example in later chapters.

The cement data have high levels of dependencies (correlations) among the predictor variables as is typical of most problems where a regression analysis might be appropriate. If all the regressor variables are mutually orthogonal (uncorrelated) then analytical considerations are more simple. Orthognality arises in controlled experiments where the factors and levels are *designed* to be orthogonal. In observational studies, there is often a high probability that some of the regressor variables will be mutually quite dependent. Rigorous experimental methods were just being developed during the time these data were taken (about 1930). Had such design methods been widely available and the importance of replication understood, then it would have been possible to break the unwanted correlations among the x variables and establish cause and effect if that was a goal.

2.3.2 *Models of Bovine TB Transmission in Ferrets*

Caley and Hone's (2002) models for disease transmission dealt with the age-specific force of infection, $\lambda(a)$ for various age classes and the age-specific disease prevalence model with ($\alpha > 0$) and without ($\alpha = 0$) disease-induced mortality. Their model for H_1 without disease-induced mortality was

$$1-\mathrm{e}^{-\lambda\alpha},$$

where $\alpha \le s$ and s is the suckling period. The corresponding model for H_1 with disease-induced mortality was

$$\frac{\lambda\left(1-\mathrm{e}^{(\alpha-\lambda)a}\right)}{\lambda-\alpha\,\mathrm{e}^{(\alpha-\lambda)a}}.$$

Simple graphs of their hazard rates help in understanding the models derived. They introduced a guarantee parameter ($\mathcal{I}$) for the period when ferrets were not exposed to infection (Fig. 2.2). If $a > s$, then the corresponding hazard models are

$$1-\mathrm{e}^{-\lambda s},$$

where s is the suckling period and no disease-induced mortality, and

$$\frac{\lambda(1-\mathrm{e}^{(\alpha-\lambda)a})}{\lambda-\alpha\,\mathrm{e}^{(\alpha-\lambda)a}}\mathrm{e}^{-\alpha(a-s)}$$

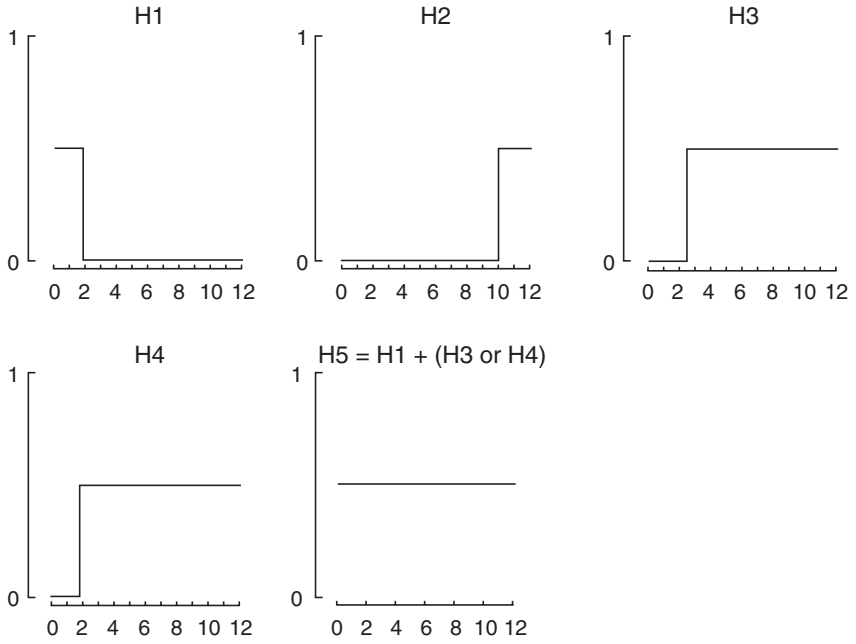

FIG. 2.2. Constant hazard functions used by Caley and Hone (2002) in modeling hypotheses concerning tuberculosis transmission in feral ferrets in New Zealand.

with disease-induced mortality. In addition, all models had a gender effect and a site effect (data were collected at seven sites). The hazard models become tedious (see their Table 2.1) and they then defined p_i to be the modeled probability of infection. A binomial likelihood function was then used where the data were y = the number of infected individuals from a total of n_i in each gender class (see data in Table 2.2). Thus, $\hat{p}_i = y_i / n_i$ as an estimate of $p = E(y_i)/n_i$. Several bounds and constraints were placed on parameter values during the optimization; additional details are given by Caley and Hone (2002). Clearly, a great deal of effort was made to derive models that accurately portrayed the hypotheses about disease transmission and the force of infection (λ) as functions of age, gender, guarantee time, and disease-induced mortality.

2.4 Additional Examples of Modeling

The first example provides some details on models and how the models might help in developing interesting science hypotheses. This is followed by further considerations in the *Exercises* section. The other two examples are more typical where the task is merely to well represent the science hypotheses by models.

2.4.1 Modeling Beak Lengths

Beak size bimodality in Darwin's finches (*Geospiza fortis*) on the island of Santa Cruz, Galapagos, Ecuador has been of interest since the early 1960s. Hendry et al. (2006) provide some background and analysis results on this set of evolutionary issues. Here we will take a hypothetical view of the data and general science question and provide alternative approaches to provide insights into hypothesizing and modeling. This example will use just beak length, while Hendry et al. (2006) performed a principal components analysis on several measurements to estimate beak "size." I will not address these real world complexities here as I want to focus on a different way to approach the evolutionary questions of interest. This approach is not claimed to be better in any way; only different to give the reader a feeling for both hypothesizing and modeling. Interested readers are encouraged to read Hendry et al. (2006) for their results with the real data.

Beak length data were collected on 1,755 birds during 1964–2005 at Academy Bay, adjacent to the town of Puerto Ayora. Histograms of the measurements suggested bimodality in the early years; however, this bimodality was lost in concert with marked increases in human population density and activity over time. This observation led to hypotheses about evolutionary forces promoting bimodality and driving adaptive radiation into multiple species over time. Perhaps the increased human disturbance blocked or at least hampered the radiation and bimodality in recent years at this site. While this extension of the problem is only to illustrate some principles, it will follow some aspects of the real situation described by Hendry et al. (2006).

Before proceeding, it is interesting to note a confirmatory aspect of this study. There are many variables that have changed on this island over the past 40–50 years. A descriptive approach might have taken measurements on many variables and asked which is the better predictor or which variables have the highest adjusted R^2 value? This is a "shot gun approach" and exposes the investigator to a high probability of finding spurious effects. The confirmatory approach asks a more specific question (is human disturbance associated with evolutionary changes in bill length?), after trying to think hard about the issue.

We first hypothesize that the bimodality observed in the histograms (per year sample sizes were roughly 100) was largely an artifact. Perhaps another bin size for the histograms would not show any pronounced bimodality. Thus, we begin with the hypothesis (H_1) that the sample data were taken from a unimodal population (this model might be of particular use in the analysis of data for the later years). Beak lengths cannot be negative; so I will use the gamma model instead of the more usual normal model. The gamma distribution (denote this first, unimodal, model as g_1) is

$$g_1(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}.$$

This model (a PDF) has two parameters, α and β, and x is the beak length. Note the expanded model notation to make it clear that the model is a function of the data x. This complicated looking model has a nice simple form (Fig. 2.3) and seems adequate as a mathematical representation of the measurement data on beak lengths (assuming unimodality).

This distribution is useful in that its mean value is estimated as $\hat{\alpha}\hat{\beta}$ and the variance is estimated as $\hat{\alpha}\hat{\beta}^2$. The shape of the distribution changes depending on the values of α and β; thus, the gamma distribution, like many statistical distributions, is a family of curves. Note too, in this type of modeling there is no response variable, instead it is the distribution of bill lengths (x) that is being modeled.

Considering the apparent observed bimodality, one might consider a *mixture* of two gamma distributions as a second model. This hypothesis assumes there is a small-beaked phenotype with some variability across individuals around a mean. Similarly, a large-beaked phenotype has some variability across individuals around a different (larger) mean. Thus, it is quite possible that an individual from the small-beaked phenotype might have a longer beak than an individual from a large-beaked phenotype. The data are hypothesized to be an unknown mixture of the two (perhaps highly variable) phenotypes. These considerations lead to a model (g_2) for the second hypothesis, H_2 (Fig. 2.4):

$$g_2 = \pi(g_s) + (1-\pi)(g_b),$$

where the parameter π is the "mixture coefficient" and g_s and g_b are gamma distributions for small (s)- and big (b)-beaked individuals, respectively. Here, $0 \le \pi \le 1$ and is the proportion of the population that have small beaks. For example, if 17% of the individuals in a given year were from a population of small-beaked phenotypes, then

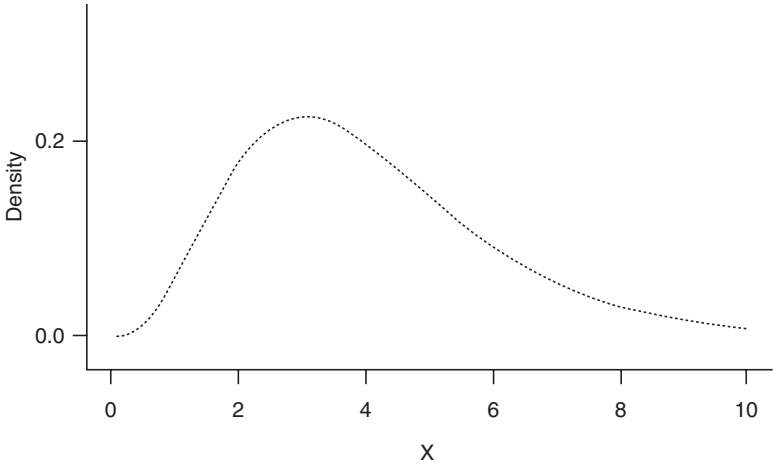

FIG. 2.3. The hypothesis of unimodality in finch bill lengths is represented as a gamma distribution.

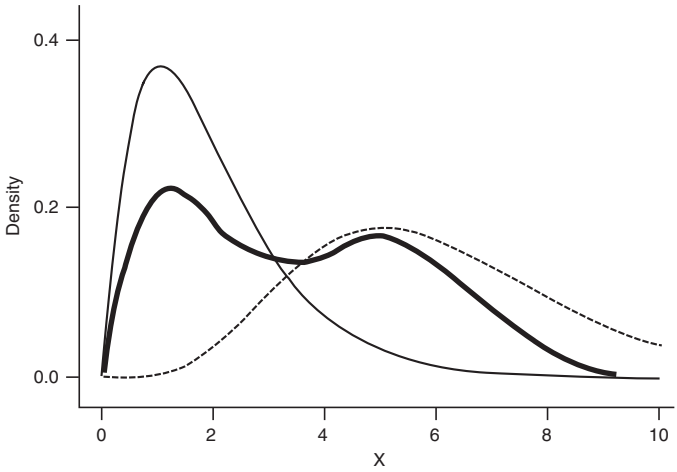

FIG. 2.4. The model representing the hypothesis that finch bill lengths arise from an unknown mixture of two phenotypes, each phenotype is modeled as a gamma distribution.

$$g_2 = 0.17(g_s) + (0.83)(g_b).$$

Of course, the two gamma distributions above would each have parameters α and β to specify the exact shape of the distributions. This model has five parameters: π, α, and β for the small-beaked animals and an α and β for the big-beaked animals. If we have a way to measure the strength of evidence for these two models we could answer questions about unimodality vs. bimodality: i.e., compare models g_1 vs. g_2.

Now we hypothesize (H_3), a linear change in bimodality over years and this is easily done by adding a submodel on the mixture coefficient π. We take model 2 and extend it to obtain a model that allows bimodality to change (drift) over years:

$$g_3 = \pi(g_s) + (1-\pi)(g_b)$$

and replace the parameter π (in two places) with the submodel

$$\pi = \beta_0 + \beta_1(T),$$

where T is the year of the study. The parameter π no longer appears in the model as it is replaced by the submodel that allows the mixture to be a function of year. Carrying out this substitution,

$$g_3 = \big(\beta_0 + \beta_1(T)\big)\cdot(g_s) + \big(1-(\beta_0 + \beta_1(T))\big)\cdot(g_b).$$

This model has six parameters: β_0, β_1, α, and β for the small-beaked animals, and an α and β for the big-beaked animals (π has been deleted and the two β parameters added). Hendry et al. (2006) hypothesized that bimodality decreased after the first few years (T), thus we expect $\hat{\beta}_1$ to be negative in this example. This model gets directly at the main evolutionary hypotheses; this is the role of these models.

We can hypothesize still other plausible alternatives, as Chamberlin would have urged. The bimodality was hypothesized to change over years (that is H_3) but perhaps caused or at least influenced by human population density over time (T) and associated disturbance (denote this environmental covariate as X_1 not to be confused with bill length, x). This covariate was measured; so we have the model (g_4) to represent the fourth hypothesis (H_4):

$$g_4 = \pi(g_s) + (1-\pi)(g_b).$$

Now replace the mixture coefficient π with a similar submodel, but with the human covariate as

$$\pi = \beta_0 + \beta_1(X_1).$$

In a sense, g_3 asked *what?* while g_4 begins to ask *why?*

We now turn our attention to a supposed covariate dealing with yearly precipitation (denote this as X_2). We will assume this variable has been measured and we will let it enter the analysis as binary: 1 for heavy precipitation and 0 for virtually no rainfall (the usual case). One can already see the pattern here as we will hypothesize (H_5) that the bimodality is influenced by (only) precipitation over the years of the study. Its associated model is

$$g_5 = \pi(g_s) + (1-\pi)(g_b),$$

where $\pi = \beta_0 + \beta_1(X_2)$.

An astute biologist then hypothesizes (H_6) that bimodality is influenced by both human activity (X_1) and precipitation (X_2), leading to an expanded submodel for π:

$$g_6 = \pi(g_s) + (1-\pi)(g_b),$$
$$\pi = \beta_0 + \beta_1(X_1) + \beta_2(X_2).$$

Finally, investigators hypothesize (H_7) to reflect interest in an interaction term in the submodel for π as

$$g_7 = \pi(g_s) + (1-\pi)(g_b),$$
$$\pi = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \beta_3(X_1 * X_2).$$

This model has eight parameters: β_0, β_1, β_2, β_3, α, and β for the small-beaked animals, and an α and β for the big-beaked animals. [I hope it is clear that the β parameters differ from model to model; i.e., the value for the MLE for β_1 and the interpretation differ by model.]

Given the sample of 1,755 beak measurements and the seven models of the seven science hypotheses, one could estimate the model parameters using maximum likelihood (see Appendix A) and proceed with a formal analysis of the evidence for each of the seven. Note that there is a nice one-to-one mapping of each hypothesis with its model. Of course, each submodel could have been hypothesized to be quadratic (or even cubic), but additional β parameters would be needed to chase these potential nonlinearities. This example attempts to show how hypothesizing and modeling can have catalytic effects. We will see this example again in the subsequent chapters.

2.4.2 *Modeling Dose Response in Flour Beetles*

Young and Young (1998:510–514) give as an example (originally from Bliss 1935) of modeling acute mortality of flour beetles (*Tribolium confusu*) caused by an experimental five-hour exposure to gaseous carbon disulfide (CS_2). The data are summarized in Table 2.3. The sample size is the 471 beetles in the dose–response experiment. One can see from Table 2.3 that the observed mortality rate increased with dosage. It is typical to fit a parametric model to effectively smooth such data, hence to get a simple estimated dose–response curve and confidence bounds, and to allow predictions (perhaps even outside the dose levels used in the experiment (i.e., extrapolation)).

A generalized linear models approach may easily, and appropriately, be used to model the probability of mortality, π_i, as a function of dose level x_i. The likelihood function for the data for a single dose is assumed to be binomial and is proportional to

$$\mathcal{L}(\pi \mid n \text{ and } y, \text{binomial}) \propto \pi^y (1-\pi)^{n-y}.$$

This notation (above) is read – the likelihood of the unknown mortality parameter π, *given* the data (the n and y) and the binomial model. The likelihood function would be different with different data or when using a model other than the binomial. Use of the binomial model brings certain assumptions with

TABLE 2.3. Flour beetle mortality at eight dose levels of CS_2 (from Young and Young 1998).

	Number of beetles		
Dose (mg/L)	Tested	Killed	Observed mortality rate
49.06	49	6	0.12
52.99	60	13	0.22
56.91	62	18	0.29
60.84	56	28	0.50
64.76	63	52	0.83
68.69	59	53	0.90
72.61	62	61	0.98
76.54	60	60	1.00

it, such as independence). Note, as is always the case, likelihoods are products of probabilities and functions of only the unknown parameters; everything else is known (i.e., given). Shorthand notation includes $\mathcal{L}(\pi|\text{data})$ or just $\mathcal{L}$ if the context is clear. The symbol "$\propto$" means "proportional to" because a constant term (the binomial coefficient), independent of the model parameters, has been omitted (Appendix A).

The flour beetles were dosed at eight levels and the likelihood for the entire data set is merely a product of the eight binomial likelihoods (given the usual assumption of independence, which seems quite reasonable here):

$$\mathcal{L}\left(\pi_i \mid n_i \text{ and } y_i, \text{binomial}\right) \propto \prod_{i=1}^{8} \left(\pi_i\right)^{y_i} \left(1-\pi_i\right)^{n_i - y_i}$$

or just the shorthand

$$\mathcal{L}\left(\pi_i \mid \text{data}\right) \propto \text{ or } L \propto \prod_{i=1}^{8} \left(\pi_i\right)^{y_i} \left(1-\pi_i\right)^{n_i - y_i}.$$

This likelihood sets up a model of the unknown mortality probabilities, but they do not depend on dose. Thus, we can hypothesize some monotonic parametric submodels involving dose $\pi_i \equiv \pi(x_i)$. I will denote dose at level i simply as x_i and constrain the probability of mortality (π) to be within 0–1.

In the context of generalized linear models, there must be a nonlinear transformation (i.e., link function) of $\pi(x)$ to give a linear structural model in the parameters. There are several commonly used forms for such a link-function based linear model but no single model form that is theoretically the correct, let alone true, one. We consider three commonly used generalized linear models and associated link functions: logistic, hazard, and probit. Each of these models has two unknown parameters that may be estimated from the data using ML. The logistic model form is

$$\pi(x) = \frac{1}{1+\mathrm{e}^{-(\alpha+\beta x)}}$$

with link function

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log \mathrm{it}(\pi(x)) = \alpha + \beta x.$$

The hazard function and the associated complementary log–log link function are

$$\pi(x) = 1 - \exp\{-\mathrm{e}^{(\alpha+\beta x)}\}$$

and

$$\log[-\log(1-\pi(x))] = \mathrm{clo\,g}\log(\pi(x)) = \alpha + \beta x.$$

The cumulative normal model and associated probit link are

$$\pi(x) = \int_{-\infty}^{\alpha+\beta x} \left[\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(1/2)z^2} \right] \mathrm{d}z \equiv \Phi(\alpha + \beta x)$$

and

$$\Phi^{-1}(\pi(x)) = \text{probit}(\pi(x)) = \alpha + \beta x.$$

Here, $\phi(\bullet)$ denotes the standard normal cumulative probability distribution, which does not exist in closed form.

In each of the three cases above, the model is sigmoidal, bounded by 0 and 1, and has two parameters. These are little more than descriptive models; i.e., they have about the "right" shape and have been useful in this class of experiments since the 1930s. The link functions let the investigator "think" of the models as simple regressions, $\alpha + \beta x$ and this is a useful construct.

Substituting the logistic model for dose level into the likelihood for a product of binomials gives

$$\mathcal{L} \propto \prod_{i=1}^{8} \left(\frac{1}{1+\mathrm{e}^{-(\alpha+\beta x_i)}} \right)^{y_i} \cdot \left(1 - \frac{1}{1+\mathrm{e}^{-(\alpha+\beta x_i)}} \right)^{n_i - y_i},$$

where x_i is the dose level and the likelihood ($\mathcal{L}$ above) is formally $\mathcal{L}(\alpha,\beta|\text{data})$. Thus ML can be used to get the MLEs $\hat{\alpha}$ and $\hat{\beta}$(the probabilities of mortality are removed and they are replaced by a simple function of dose level, x_i). This particular example happens to be logistic regression and can be done easily in software packages (e.g., SAS Institute 2004). This is another example where one might start with a simple model, such as the binomial model here. Assuming independence (Sect. 6.1 and Appendix A), one can take the product of all eight binomials as the likelihood. Then, adding submodels in place of one or more model parameters can bring tremendous flexibility and realism to the modeling process.

At some early point we must ask if a model fits the data in a reasonable way. A simple Pearson observed vs. expected chi-square comparison often suffices as a goodness-of-fit (GOF) assessment:

$$\chi^2 = \Sigma \frac{(O_j - \hat{E}_j)^2}{\hat{E}_j}$$

where O_j is the observed values and $\hat{E}_j$ is the estimated expected values. These test statistics each have six degrees of freedom (=8 − 2, as each model has two estimated parameters). The chi-square statistic is

$$\chi^2 = \sum_{i=1}^{8} \frac{\left(y_i - n_i \hat{\pi}_i\right)^2}{n_i \hat{\pi}_i \left(1 - \hat{\pi}_i\right)}.$$

Goodness-of-fit results are

Model	χ^2
cloglog	3.49
probit	7.06
logit	7.65

indicating a good fit for each of the three models. Normally GOF is assessed only for the global model; in this case there is no such model, but three non-nested competitors all with $K = 2$ unknown model parameters (see Appendix A). If the response variable is continuous, there are a large number of standard diagnostics and procedures to analyze residuals; these are widely available in computer software.

A key feature of this beetle mortality example is causality. The experimentally applied dose *caused* the observed mortality. By the design we can establish *a priori* that (1) the only predictor needed, or useful, is dose and (2) monotonicity of expected response should be imposed (i.e., the higher the dose, the higher the probability of death). The issue about a model is thus reduced to one of an appropriate functional form, hence, in a generalized linear models framework, to what is the appropriate link function. However, as a result, we have no global model, but rather several (three were used) alternatives for a best causal-predictive model (many observational studies lack a global model).

2.4.3 *Modeling Enzyme Kinetics*

Over many years, a series of models have been developed for understanding enzyme inhibition (Nelder 1991; Brush 1965, 1966). This field has matured and I will give some general models that have found use in this issue. We will review four models, each representing a hypothesis concerning the rate of enzyme-mediated reaction (R). There are only two predictor variables: S = substrate concentration and I = inhibiting substance. Four hypotheses are represented by the following models:

H_1 noncompetitive (general) model — Parameters

$$R = \frac{\beta_1 \cdot S}{\beta_2(1-\beta_3 \cdot I) + S(1+\beta_4 \cdot I)} \qquad \{\beta_1,\beta_2,\beta_3,\beta_4,\sigma^2\} = 5$$

H_2 Michaelis–Menten model

$$R = \frac{\beta_1 \cdot S}{\beta_2 + S} \qquad \{\beta_1,\beta_2,0,0,\sigma^2\} = 3$$

H_3 competition model

$$R = \frac{\beta_1 \cdot S}{\beta_2 (1 - \beta_3 \cdot I) + S} \qquad \{\beta_1, \beta_2, \beta_3, 0, \sigma^2\} = 4$$

H_4 uncompetitive model

$$R = \frac{\beta_1 \cdot S}{\beta_2 + S(1 + \beta_4 \cdot I)} \qquad \{\beta_1, \beta_2, 0, \beta_4, \sigma^2\} = 4$$

Two of the model parameters have scientific interpretations: β_1 is the maximum reaction rate and β_2 is the half saturation level. Parameters β_3 and β_4 are called inhibition kinetic values. Here each hypothesis has been represented by its associated model and we can speak of hypothesis *i* or model *i* synonymously. The parameters can be estimated using ML methods and inferences made. Critically, we would like measures of the evidence for each of the four hypotheses in the set: "What is the empirical support for hypothesis *i* vs. *j*?" A set of models such as this does not arise overnight; instead, these models are the result of much effort in the laboratory and much analytical thought. Model building should take full advantage of past research.

There are many model based studies in human medicine but my opinion is that often only a single hypothesis and its model are the focus of the study. Clyde (2000) and Remontet et al. (2006) provide examples of multiple models and model selection in this important area.

2.5 Data Dredging

Data dredging (also called *post hoc* data analysis) begins after the planned (*a priori*) analysis and after inspecting those results. Data dredging should generally be minimized or avoided, except in (1) the early stages of exploratory work or (2) *after* a more confirmatory analysis has been completed. In this latter case, the investigator should fully admit to the process that lead to the *post hoc* results and should treat the results much more cautiously than those found under the initial, *a priori* approach. One approach in *post hoc* analyses is to start with the best model (from the *a priori* results) and expand around it. When done carefully, we encourage people to explore their data beyond the important *a priori* phase. Still, *post hoc* results are like skating on thin ice – lots of risks of getting in trouble (i.e., finding effects that are spurious because noise is being modeled as structure).

I recommend a substantial, deliberate effort to get the *a priori* thinking and models in place and try to obtain more confirmatory results; *then* explore the *post hoc* issues that often arise after seeing the more confirmatory results. Data dredging activities form a continuum, ranging from fairly trivial (venial) to the grievous (mortal). There is often a fine line between dredging and not dredging; my advice is to stay well toward the *a priori* end of the continuum and thus achieve a more confirmatory result. One can always do *post hoc*

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

Grievous data dredging is endemic in the applied literature and still frequently taught or implied in statistics courses without the needed caveats concerning the attendant inferential problems. Rampant rummaging through the data looking for patterns and then "testing" them would be called, in any other human endeavor, *cheating*.

Running all possible models is a thoughtless approach and runs the high risk of finding effects that are, in fact, spurious if only a single model is chosen for inference. If prediction is the objective, model averaging is useful and estimates of precision should include model selection uncertainty; these are subjects to be addressed in later sections of this book. Even in this case, surely one can often rule out many models on *a priori* grounds (e.g., the cement hardening data). There are recent papers in major journals that provide the results of analyses where well over a million models have been run with sample sizes <100. I suspect nearly every result was actually spurious in such cases. Running all possible models is usually a signal of an unthinking science approach.

2.6 The Effect of a Flood on European Dippers: Modeling Contrasts

Lebreton et al. (1992) provided a small set of capture–recapture data on the European Dipper (*Cinclus cinclus*). This is a small bird that spends its life along small streams; the data come from eastern France and were collected by Marzolin. The study took place over seven years; thus there are six survival intervals. A flood took place toward the end of the second survival interval and continued into the beginning of the third survival interval. The simple science question asked if survival probability was lower in the two flood years. Note that causation (the flood caused lowered survival) cannot be addressed here as this is an observational study, not a strict experiment.

Some notation is needed; let φ be the time-averaged annual survival probability while φ_{f} and φ_{nf} be the time-averaged annual survival probabilities for flood and nonflood years, respectively. Specifically, φ is the conditional probability that a dipper survives the annual interval and stays on the study area, given it is alive at the beginning of the interval. Finally, we denote the time-averaged probability of capture or recapture as p.

2.6.1 Traditional Null Hypothesis Testing

Standard practice would be to define a null and alternative hypothesis and their corresponding models. The null hypothesis (H_0) would be that there is

no effect (exactly no effect) of the flood on annual survival probability, while the alternative hypothesis (H_a) would be that the flood did have an effect on annual survival probability. So, we have two models representing the null and alternative hypothesis, respectively:

H_0: $\{\varphi,p\}$ with two unknown parameters
H_a: $\{\varphi_f,\varphi_{nf},p\}$ with three unknown parameters

The null model is nested within the more general alternative model and this fact allows standard "tests" to be computed to address the issue of a flood effect on annual survival probabilities. This test is done by testing (only) the null hypothesis; the alternative is *not* the subject of the test. If the null is rejected, then, *by default*, the alternative is said to be supported. The alternative hypothesis (the one the investigator usually believes) is never tested.

2.6.2 Information-Theoretic Approach

The information-theoretic approach would begin with the same two hypotheses, $\{\varphi,p\}$ and $\{\varphi_f, \varphi_{nf},p\}$, claiming that these models are only simple approximations to the complex reality. There is no need that the models are nested (they happen to be in this case). The information-theoretic approach asks for measures of relative support (i.e., from the data, empirical) for the two hypotheses. It is not alleged that hypothesis $\{\varphi,p\}$ is exactly true; rather it is a hypothesis and a model that are approximations.

In this early example, perhaps relatively little thought went into the hypotheses to be included in the set – two hypotheses seem "obvious." A little more thought suggests that other hypotheses could be examined (as Bacon and Chamberlin would have wanted):

- Was there a survival effect just the first year of the flood $\{\varphi_{f1},\varphi_{nf},p\}$?
- Or just the second year of the flood $\{\varphi_{f2},\varphi_{nf},p\}$?
- Or was the recapture probability (p) also effected by the flood $\{\varphi_f,\varphi_{nf},p_f,p_{nf}\}$?
- Or even $\{\varphi,p_f,p_{nf}\}$, where survival was not impacted, but the recapture probabilities were?

Note that few of the models above are nested; thus each model must be tested against the null and this raises the multiple testing problem, a scourge of null hypothesis testing. Traditional tests do not allow much evidence about the relative merit of the four hypotheses/models above.

Thinking hard about hypotheses to be evaluated before data analysis nearly always has its clear rewards. In this simple example, the addition of four more hypotheses was not particularly "heavy mental lifting" but in more challenging problems considerable thought is usually required. We must all do more to encourage a culture of hard thinking and rigor in scientific work. A premium

must be placed on thinking, innovation, and creativity – do not expect the computer to tell us what is "important."

Simple problems such as the dipper problem can be effectively addressed with the methods developed in this text; just because the problem is simple does not mean one must use null hypothesis testing methods.

2.7 Remarks

Romesburg (2002) wrote a fascinating book about thinking and the creative spirit; I have found this very useful and recommend it.

Chatfield (1995a,b) provides very good guidelines concerning statistical practice. Gotelli and Ellison (2004) provide sage advice on data handling and archiving. Manly's (1992) book covers both sampling and design issues and is easy reading.

Chamberlin's paper is well worth reading after more than 100 years. How many papers in *Science* have been reprinted in the same journal (as was Chamberlain's in 1890 and in 1965)?

Fisher first published on his likelihood approach when he was a third year undergraduate (1912) and a very much extended account in 1922. Likelihood is among the great achievements in statistics (like aspirin in medicine); it is the backbone of statistical thinking, including Bayesian approaches. It might be noted that the Fisher information matrix addresses *precision* (a measure of repeatability) when translated into the covariance matrix, rather than strictly information. Of course, precision is tied to "information." The first book (Edwards 1976) on likelihood was written well after Fisher's fundamental paper on the subject in 1922; this was followed by an expanded treatment in 1992. It is fitting that Edwards was Fisher's last Ph. D. student. Oddly, there are still relatively few books on the subject (good examples include Azzalini 1996; Royall 1997; Severini 2000; Pawitan 2001); like the ubiquitous "delta method"–everyone is supposed to (somehow) know it!

Draper and Smith (1981) provide a review of results found by others that have analyzed the cement hardening data (also see Hald 1952 and Hand 1994). Hendry et al. (2006) give an analysis of the actual data on beak size in Darwin's finches; the hypothetical example here takes their work in a conceptually different modeling direction. Additional results on bovine tuberculosis in feral ferrets in New Zealand are provided by Caley and Hone (2005).

Many papers exist on modeling but there is a clear need for a nice book synthesizing the literature and providing effective examples. Levins (1966), Leamer (1978), Gilchrist (1984), Lehman (1990), Starfield et al. (1990, 1991), Cox (1990, 1995), O'Connor and Spotila (1992), Scheiner and Gurevitch (1993), and Lunneborg (1994). Chatfield (1995a,b, 1991), Nichols (2001), and Shenk and Franklin (2001), and Zuur (2007) offer good introductions into

the statistical modeling literature. White and Lubow (2002) provide examples of modeling data from differing sources.

Much of statistical theory is based on an assumption about so-called independence and this is often compromised with data in the life sciences. What is required, in general, is a correct likelihood for the data that reflects any dependence. There is a simple way to handle some lack of independence in making inferences (Sect. 6.1). An easy reading paper on spurious effects and how to minimize these is Anderson et al. (2001a). Inferential problems when using convenience sampling are outlined by Anderson (2001), but see also Hairston (1989) and Eberhardt and Thomas (1991).

2.8 Exercises

1. Reread the paper by Caley and Hone (2002).
   a. They demonstrated that estimating the force of infection (λ) from age-prevalence data is possible and assists in discriminating between alternative hypotheses about routes of disease transmission. Discuss this finding and compare it with similar studies of disease transmission in humans.
   b. Their hypothesis concerning dietary-related transmission from the age of weaning had the best empirical support. Think hard about this and ask if there are logical next steps in understanding the transmission issue. For example, since there was a debate or controversy over this whole issue, what might you want as an opponent?
   c. What would your value judgment be concerning inductive inferences from their sample data to the five populations of ferrets in New Zealand?
2. For those readers with an advanced understanding of mathematical statistics, what worries might you have about getting MLEs of the parameters in the seven models of beak lengths in Darwin's finches? What might be done to avoid problems here?
3. What is "wrong" in merely presenting the eight estimates of flour beetle mortality probability as a function of dose level, either in a table or a simple graph? Why go through the modeling and reparameterization (e.g., the substitution of a submodel involving the parameters β_0 and β_1 for π)? What is the principle here and what are the advantages? (Advanced question).
4. Can you think of any model in the life sciences that is strictly true? What about the physical sciences? Or medicine? Or economics? If possible, ask people in those disciplines for examples of exactly true models in their field. How would a person know with certainty that the true model was in the set, but not know which one is was? Lastly, how do we *know* a model is exactly true? Can you imagine methods that would allow one to determine (e.g., test for) the exact truth of a model? [Probably not a good Ph.D.

project.] This would be a case where the form of the true model would be known, but not its parameters. It seems a shame that true models do not come with their true parameters, making estimation unneeded!

5. Few editors, associate editors, and reviewers seem to be aware of the inferential issues with unadulterated data dredging. They seem to believe that all analysis results are created equal and it makes no difference if the hypothesis was posed before or after data analysis. Discuss this issue. How can this issue be improved so our science moves ahead more rapidly?
6. Linhart and Zucchini (1986) analyzed data on weekly storm events at a botanical garden in Durban, South Africa. They had data over 47 consecutive years and were interested in prediction of weekly storm events (i.e., $i = 1, 2, \ldots, 52$ weeks). They knew that an estimator of the probability of a storm in week i was $\hat{p}_i = y_i / 47$, where y_i is the number of storms in week i. Thus, they computed the binomial estimator (an MLE) for all 52 weeks. Critique this approach. What is "wrong" here?