Chapter 15

Hierarchical Models

The topic of this chapter is models in which the controlling parameters are taken to be a random component in addition to the random model component that describes the distribution of response random variables. The name hierarchical models stems from a hierarchy of parameter values; the data model for response variables contain parameters that themselves have distributions that depend on additional parameters. Just as the underlying basis of Bayesian analysis does not rely on treating parameters as random variables, so to does treating parameters as random variables not rely on taking a Bayesian approach to estimation and inference. That said, it is perhaps natural to consider a Bayesian analysis for many hierarchical models because Bayesian methods extend more easily to such models than do most frequentist approaches to analysis.

15.1 Mixed Models as Hierarchical Models

Consider a problem formulated in terms of groups of random variables $Y_{i,j}$ where $j = 1, \ldots, n_i$ indexes observation within group and $i = 1, \ldots, n$ indexes group. We have available one type of covariate $x_{i,j}$ and would like to relate the responses to the covariate values. A linear random effects model might be formulated for this situation as

$$Y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \tau\delta_i + \sigma\, \epsilon_{i,j}, \tag{15.1}$$

where $\delta_i \sim iid$ N$(0, 1)$ and $\epsilon_{i,j} \sim iid$ N$(0, 1)$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. Here, the *conditional* model given δ_i has expected values $E(Y_{i,j}|\delta_i) = \beta_0 + \tau\delta_i + \beta_1 x_{i,j}$ and variances $\mathrm{var}(Y_{i,j}|\delta_i) = \sigma^2$. The *marginal model* has expected values $E(Y_{i,j}) = \beta_0 + \beta_1 x_{i,j}$, variances $\mathrm{var}(Y_{i,j}) = \tau^2 + \sigma^2$ and covariances $\mathrm{cov}(Y_{i,j}, Y_{i,k}) = \tau^2$. The systematic model component of the conditional model describes parallel regression lines and model (15.1) is sometimes called a *random intercept* model. It may be written directly in that way as

$$Y_{i,j} = b_{0,i} + \beta_1 x_{i,j} + \sigma\epsilon_{i,j}, \tag{15.2}$$

where $b_{0,i} \sim iid$ N(β_0, τ^2) and $\epsilon_{i,j} \sim iid$ N$(0, \sigma^2)$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. Both the conditional and marginal models corresponding to (15.2 have the same expected values, variances and covariances as those of (15.1). In fact, the models are equivalent.

While models (15.1) and (15.2) are the same model, the ways they are written reflect a difference in how the stochastic component describing random intercepts is being conceptualized. Written as (15.1), the random effect of group can be easily conceptualized as an additional error term in a simple linear regression model. This error term is shared by all observations within a group. Written as (15.2) the random effect of group is more easily conceptualized as a random data model parameter or perhaps a better statement would be as a random variable that plays the role of a parameter in the data model. This distinction is actually more than trivial detail, because it identifies two approaches for extending the idea of random effects to more complex models. One approach, embodied by (15.1) is to consider models with multiple stochastic components as arising from variance components connected with nested data structures. This approach leads naturally to what are called *mixed effects* models. The entire book by Bryk and Raudenbusch (1992) is based on this idea. The other approach, embodied by (15.2), is to consider the parameters in a data model as representing one manifestation of a scientific mechanism, which leads naturally to models with hierarchies of parameters, namely *hierarchical models.* The hierarchical viewpoint is in concert with the discussion of statistical modeling contained in Part 1 of this book. A scientific mechanism or phenomenon of interest is represented through the parameters of a statistical model (the process of statistical abstraction). For many such mechanisms, however, it is not possible to divorce their effect from all other factors that are active at the time of observation. Thus, what we see is not caused by a mechanism that is the same in every situation, but by a particular manifestation of the mechanism. In another place or time, that same mechanism may manifest itself in a somewhat different way. The distribution of random variables that play the role of parameters in a data model describes the relative frequencies with which the mechanism manifests itself over a class of situations.

Example 15.1

Consider again the relation between Cd concentration and length in Yellow Perch from Little Rock Lake, as discussed in Chapter 8. Would we expect this same regression equation (that is, with the same parameter values) to also describe the relation of Cd concentration to length in Yellow Perch from Lake Puckaway (a lake somewhat larger than Little Rock Lake, and located in south-central rather than north-central Wisconsin)? Most likely we would not be that naive. We might believe, hope, or wish to investigate whether the same model *structure* (i.e., inverse Gaussian random component with log link) is adequate in both situations, but it would be unrealistic to assume that the same parameter values would apply. In effect, given that the relation between responses (Cd concentration) and covariate (length) does reflect a meaningful mechanism (bioaccumulation of Cd), the two lakes, Little Rock in the north and Puckaway in the south, represent two different *manifestations* of that mechanism. If our model form is adequate to describe the mechanism over a range of situations, differences in the parameter values reflect the variability in the way the mechanism is manifested under different circumstances. Now suppose that we were able to obtain observations from a variety of particular manifestations of the mechanism in, for example, k different lakes (a random sample of lakes would be good here). Then, we might take each lake as having its own regression, and model the parameters of those regressions as coming from some distribution. The mechanism we are trying to model is then embodied in the distribution of parameters, not necessarily a marginal model. It will likely be true that we need the joint marginal distribution of all our observable random variables in order to estimate that distribution, but the form of the marginal model itself may be of little concern otherwise.

This chapter presents the hierarchical viewpoint just described. The phrases *random parameters* or *random data model parameters* will be used rather than the more cumbersome *random variables that play the role of data model parameters.*

15.2 Basic Mixture Models

We will first consider situations that involve groups of independent response variables, for which each variable has its own distribution. Such models are useful in comparison of groups. These models have sometimes been considered ways to cope with what is called *overdispersion*, but we present them here in the context of hierarchical models. In addition, we will describe these models as they would be formulated for a single group; this is adequate for the comparison of groups using either frequentist or Bayesian methods of analysis. Note that the models considered here are sometimes called general mixture models to distinguish them from the finite mixture models of Chapter 13.

15.2.1 Formulation

Consider a set response variables $\{Y_i : i = 1, \ldots, n\}$, assumed to be independent given a corresponding set of parameters $\{\theta_i : i = 1, \ldots, n\}$. Let the density or mass functions of the Y_i be denoted as $\{f(y_i|\theta_i) : i = 1, \ldots, n\}$. This set of distributions then constitutes the data model or observation process. Now, let the parameters θ_i be *iid* random variables following a common density or mass function $g(\theta_i|\boldsymbol{\lambda})$. This is then the random parameter model or what we will call the mixing distribution. Following the previous discussion, the scientific mechanism or phenomenon of interest is now conceptualized as the mixing distribution controlled by the parameter $\boldsymbol{\lambda}$. We can write the joint data model as

$$f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\theta_i),$$

and the joint random parameter model as,

$$g(\theta_1, \ldots, \theta_n|\boldsymbol{\lambda}) = g(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} g(\theta_i|\boldsymbol{\lambda}).$$

The joint marginal distribution of the response variables is then derived as,

$$h(\boldsymbol{y}|\boldsymbol{\lambda}) = \int \ldots, \int f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\theta_1, \ldots, d\theta_n. \qquad (15.3)$$

Now, because of independence throughout this model formulation, it is generally simpler to derive $h(\boldsymbol{y}|\boldsymbol{\lambda})$ as

$$h(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} h(y_i|\boldsymbol{\lambda}),$$

where,

$$h(y_i|\boldsymbol{\lambda}) = \int f(y_i|\theta_i)\, g(\theta_i|\boldsymbol{\lambda})\, d\theta_i.$$

In the above general notation, we use the following nomenclature:

- $f(y_i|\theta_i)$ is the data model for Y_i
- $g(\theta_i|\boldsymbol{\lambda})$ is the mixing distribution
- $h(y_i|\boldsymbol{\lambda})$ is the resultant mixture of f over g
- $\log\{h(\boldsymbol{y}|\boldsymbol{\lambda}) = \sum_i \log\{h(y_i|\boldsymbol{\lambda})\}$ is the marginal or mixture log likelihood, and figures prominently in either likelihood or Bayesian analysis although, as we will see, analytic derivation of $h(\boldsymbol{y}|\boldsymbol{\lambda})$ is not always required in a Bayesian approach.

15.2.2 Common Mixture Models

There are a number of data model/random parameter model combinations that are frequently used and work out nicely from a mathematical viewpoint. There is nothing magic about these combinations, and our thinking should not be constrained by the following list. Nevertheless, these data model/random parameter model combinations do have nice mathematical properties and often seem reasonable combinations to use in practical situations.

1. Beta-Binomial. This mixture model takes, for $i = 1, \ldots, n$, $f(y_i|\theta_i)$ to be a binomial probability mass function with binomial sample size m_i, and $g(\theta_i|\boldsymbol{\lambda})$ to be beta distributions with parameter $\boldsymbol{\lambda} = (\alpha, \beta)$. The beta-binomial model was introduced by Williams (1982) in the context of studies of teratogenic effects.

2. Gamma-Poisson. Here, the data model $f(y_i|\theta_i)$ consists of conditionally independent Poisson distributions with parameters θ_i, and the mixing distributions $g(\theta_i|\boldsymbol{\lambda})$ are gamma distributions with parameter $\boldsymbol{\lambda} = (\alpha, \beta)$. This model is sometimes referred to as a negative binomial distribution, but under the development of a negative binomial as the number of binary trials preceding a given number of successes, this is only true for integer-valued gamma parameters.

3. Normal-Normal. In a normal-normal mixture model the data model distributions $f(y_i|\theta_i)$ are taken to be conditionally independent normal distributions with means θ_i and either known variances or variances considered as uninteresting *nuisance* parameters. The mixing distributions $g(\theta_i|\boldsymbol{\lambda})$ are then assumed to also be normal with parameter $\boldsymbol{\lambda} = (\mu, \tau^2)$.

4. Normal-Inverse Gamma-Normal. Here, the data model distributions

$f(y_i|\boldsymbol{\theta}_i)$ are conditionally independent normals with parameters $\boldsymbol{\theta}_i = (\mu_i, \sigma_i^2)$ where both expected values and variances are unknown. Typically, the mixing distributions are formulated as $g(\boldsymbol{\theta}_i|\boldsymbol{\lambda}) = g_1(\mu_i|\lambda, \tau^2)\, g_2(\sigma_i^2|\alpha, \beta)$ where g_1 is normal and g_2 is inverse gamma.

5. Multinomial-Dirichlet. For this model the response variables are vectors, $\boldsymbol{Y}_i = (Y_{i,1}, \ldots Y_{i,h})^T$, where $Y_{i,j}$ is a count of observations in category j from a set of m_i polycotomous trials. The data model assumes the $\boldsymbol{Y}_i$ are conditionally independent with multinomial probability mass functions $f(\boldsymbol{y}_i|\boldsymbol{\theta}_i)$ where $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_k)^T$. The mixing distribution $g(\boldsymbol{\theta}_i|\boldsymbol{\eta})$ is taken as Dirichlet with parameter $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$. Note that in this formulation, $\sum \theta_i = 1$ and this is a constraint that must be adhered to. Some would prefer to write the multinomial with only $k-1$ parameters and a bounded-sum constraint $\sum \theta_i < 1$, but this would require modification of the Dirichlet mixing distribution. It is also worthy of note that, while the multinomial vectors $\boldsymbol{Y}_i$; $i = 1, \ldots, n$ are assumed to be independent, there will be negative correlation within the components of $\boldsymbol{Y}_i$ due to the constrained sum.

The alert reader will be aware that the list just presented is very similar to what were given in Chapter 7 as pairs of conjugate data models and priors. There is, in fact a mathematical connection in that in a the integral needed to determine the denominator of (7.1) in closed form is the same as the integral needed to determine (15.3) in closed form. But **there is no statistical connection**. And one should not form the opinion that the marginal distribution of response random variables (15.3) must be able to be derived analytically in order for estimation and inference to be possible. The conditional independence of the data model combined with the *iid* structure of the mixing distribution allow numerical approximations to the marginal distribution to be used effectively.

Example 15.2

Consider a data model consisting of normal distributions with expected values $-\infty < \mu_i < \infty$ and known variance 1,

$$f(y_i|\mu_i) = \frac{1}{(2\pi)^{1/2}} \exp\left[-\frac{1}{2}(y_i - \mu_i)^2\right]; \;\; -\infty < y_i < \infty.$$

Assume the μ_i are *iid* with a common extreme value distribution, for $-\infty < \xi < \infty$ and $\theta > 0$,

$$g(\mu_i|\xi,\theta) = \exp\left(-\frac{\mu_i - \xi}{\theta}\right) \exp\left[-\exp\left(-\frac{\mu_i - \xi}{\theta}\right)\right]; \;\; -\infty < \mu_i < \infty.$$

The marginal log likelihood is,

$$\begin{aligned} \ell(\xi,\theta) &= \sum_{i=1}^{n} \log\{h(y_i|\xi,\theta)\} \\ h(y_i|\xi,\theta) &= \int_{\mathbb{R}} f(y_i|\mu_i)\, g(\mu_i|\xi,\theta)\, d\mu_i, \end{aligned} \tag{15.4}$$

and the integral in (15.4) is not tractable. For a set of observed values $y_1, \ldots, y_n$ and parameters ξ and θ, however, the log likelihood $\ell(\xi,\theta)$ requires only a set of n independent one-dimensional integrations, which is not difficult to program. As a result, an iterative optimization algorithm such as the black-box functions in `R` can be used to find maximum likelihood estimates of ξ and θ. If one desires to program derivatives as well as the log likelihood itself, one need only pass the derivatives under the integral and again numerically evaluate sets of one-dimensional integrals.

15.3 General Mixture Models

In the previous section, we restricted attention to situations in which each individual response random variable Y_i followed a distribution with a unique value of its parameter θ_i. Combined with an *iid* structure for the mixing distribution this led to the responses having *iid* marginal distributions. Mixture models can certainly be applied in more general settings involving, for example, groups of response random variables that share a value of the data model parameter. We give several examples to illustrate the possibilities.

Example 15.3

Consider a linear random effects model, written in the form of a data model with random parameters, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$,

$$Y_{i,j} = \beta_{0,i} + \beta_1 x_{i,j} + \epsilon_{i,j},$$

where $\beta_{0,i} \sim iidN(\psi, \tau^2)$ and $\epsilon_{i,j} \sim iidN(0, \sigma^2)$. Here, the likelihood has independent pieces across values of i, but each piece must be a joint across values of j for that i. In particular, the joint data model can be written as,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\beta}_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} f(y_{i,j}|\beta_{0,i}, \beta_1, \sigma^2) \\ &= \prod_{i=1}^{n} f(\boldsymbol{y}_i|\beta_{0,i}, \beta_1, \sigma^2), \end{aligned}$$

where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n_i})^T$. The mixing distribution is,

$$g(\boldsymbol{\beta}_0|\psi, \tau^2) = \prod_{i=1}^{n} g(\beta_{0,i}|\psi, \tau^2).$$

The marginal likelihood is then the $n-$dimensional integral, reduced to a product of one-dimensional integrals as,

$$\begin{aligned} h(\boldsymbol{y}|\psi, \beta_1, \tau^2, \sigma^2) &= \int_{\mathbb{R}^n} f(\boldsymbol{y}|\boldsymbol{\beta}_0, \beta_1, \sigma^2) g(\boldsymbol{\beta}_0|\psi, \tau^2)\, d\beta_{0,1} \ldots d\beta_{0,n} \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} f(\boldsymbol{y}_i|\beta_{0,i}, \beta_1, \sigma^2)\, g(\beta_{0,i}|\psi, \tau^2)\, d\beta_{0,i}. \end{aligned}$$

Example 15.4

Suppose that in the problem of relating Cd concentration to length of yellow perch, we have data available from n lakes rather than just one. We might anticipate that, as with the data of the Chapter 8 case study, the data from each lake could be described using a generalized linear model with gamma random component and log link function. We might then index lakes as $i = 1, \ldots, n$ and observations within a lake as $j = 1, \ldots, n_i$. A general mixture model for this situation might use gamma distributions parameterized by $\mu_{i,j}$ and ϕ_i as in the usual exponential dispersion family form. We could then formulate a model as,

$$\begin{aligned} Y_{i,j} &\sim \text{ indep. Ga}(\mu_{i,j}, \phi_i) \\ \log(\mu_{i,j}) &= \gamma_{0,i} + \gamma_{1,i} x_{i,j} \\ \gamma_{0,i} &\sim \text{ iid N}(\mu_0, \sigma_0^2) \\ \gamma_{1,i} &\sim \text{ iid N}(\mu_1, \sigma_1^2) \\ \phi_i &\sim \text{ iid Ga}(\lambda_1, \lambda_2) \end{aligned}$$

Here, the joint data model would be

$$f(\boldsymbol{y}|\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\phi}) = \prod_{i=1}^{n}\prod_{j=1}^{n} f(y_{i,j}|\beta_{0,i}, \beta_{1,i}, \phi_i) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|\beta_{0,i}, \beta_{1,i}, \phi_i).$$

The marginal joint would then be

$h(\boldsymbol{y}|\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \lambda_1, \lambda_2) =$

$$\prod_{i=1}^{n} \int_{\mathbb{R}^2 \times \mathbb{R}^+} f(\boldsymbol{y}_i|\gamma_{0,i}, \gamma_{1,i}, \phi_i) g_0(\gamma_{0,i}|\mu_0, \sigma_0^2) g_1(\gamma_{1,i}|\mu_1, \sigma_1^2) g_\phi(\phi_i|\lambda_1, \lambda_2)\, d\gamma_{0,i}\, d\gamma_{1,i}\, d\phi_i.$$

15.4 Case Study: Selenium in California

The Central Valley of California is a large agricultural region, but primarily because of extensive irrigation. The Central Valley was originally quite arid, and the underlying geology is that of an ancient sea bed. It is also a historical stopping ground for migratory waterfowl in what is called the Pacific Flyway, a broad corridor for waterfowl that breed in the north (e.g., Alaska and British Columbia) but winter in Mexico and Central America. When one irrigates an area heavily, over a period of years the water table rises. If that area was formed on the sedimentary material of an ancient sea bed, the underlying bedrock contains minerals and salts which become dissolved as excess irrigation water percolates down through the soil. When the water table rises to the level of the root zone of plants, the salinity kills the plants. The engineering solution to this problem is to tile agricultural fields, and drain excess irrigation water from the ground. There of course needs to be a depository for this irrigation return flow which, in the Central Valley was accomplished by construction of a great number of evaporation ponds. The original thought was that such ponds would also be ideal habitat for migrating waterfowl as they moved through the area, as well as holding the potential for benefits from human recreation. But, when salt and mineral-laden irrigation return water evaporates it leaves behind much of the salt and mineral burden, which can become toxic in high concentrations. When the evaporation ponds in the Central Valley began to yield deformed frogs (e.g., six legs, two heads but no legs) and other aquatic life, concern was raised for both the health of the ecosystem and potential implications for humans using the ponds for fishing, boating, and other recreational activities.

Attention eventually focused on Selenium (Se), a necessary trace element for life to exist, but teratogenic in high concentrations. A contentious issue, however, was whether Se was in fact causing problems in the real world, or whether it could only be shown to have an effect in controlled laboratory studies using unrealistically high exposures. A large number of field studies of the Central Valley region ensued. One such study was conducted by the U.S. Fish and Wildlife Service to examine the potential teratogenic effect of irrigation return water to aquatic life by looking at reproductive success in Mosquitofish *Gambusia spp.* (Saiki and Ogle, 1995). *Gambusia* are a small fish that form the basis of many aquatic food chains in this region, and they are also one of the few fish taxa that are viviparous (give live birth).

Now, for irrigation return water to be delivered to evaporation ponds requires the construction of what are called irrigation return flow canals. One of the larger of these in the Central Valley is called the *San Luis Drain.* In 1983 a large fish kill was observed in the San Luis Drain and Se levels at that time were extremely high. From 1983 to 1985, *Gambusia* was the only fish observed in the San Luis Drain, although previously the canal had supported populations of largemouth bass, striped bass, catfish, and bluegill, among others. A nearby area, the *Volta National Wildlife Refuge* receives no irrigation return water, and did not experience a similar fish kill.

In June, 1985, gravid female *Gambusia* were collected from both the San Luis Drain and the Volta NWR. Fish of similar length, weight, and stage of pregnancy were held in the laboratory until parturition and the number of live and stillborn young counted for each female.

The data used here consisted of observations of the total number of young for each female and the number of young born live. Thus, high proportions are an indication of good reproductive success while low proportions are an indication of poor reproductive success. The observed proportions available from the data indicate the presence of overdispersion, that is, more variability among females than if all individuals within a treatment group were conceptualized as generating identical binomial outcomes (we can find a test for this, (e.g., Snedecor and Cochran, 1967). As a result, a beta-binomial model was fit to each group.

For one group (SLD or Volta) let Y_i; $i = 1, \ldots, m$ be random variables associated with the number of live young produced by female i. Let m_i; $i = 1, \ldots, n$ be the total number of young for female i; we will consider the m_i fixed constants, although it would certainly be reasonable to also model them as random variables in a more complex structure. Given parameters θ_i; $\imath = 1, \ldots, n$, assume that the Y_i are conditionally independent with probability mass functions

$$f_i(y_i|\theta_i) = \frac{m_i!}{y_i!\,(m_i - y_i)!}\,\theta_i^{y_i}\,(1-\theta_i)^{m_i - y_i}, \tag{15.5}$$

for $y_i = 0, 1, \ldots, m_i$ and where $0 < \theta_i < 1$. Further, assume that θ_i; $i = 1, \ldots, n$ are *iid* with probability density functions,

$$g(\theta_i|\alpha,\, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta_i^{\alpha-1}\,(1-\theta_i)^{\beta-1}, \tag{15.6}$$

for $0 < \theta_i < 1$ and where $0 < \alpha$ and $0 < \beta$.

Combining the data model (15.5) and the mixture (or random parameter model) (15.6) yields the marginal pmf,

$$\begin{aligned} h(y_i|\alpha,\beta) &= \frac{m_i!}{y_i!\,(m_i-y_i)!}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\int_0^1 \theta_i^{\alpha+y_i-1}\,(1-\theta_i)^{\beta+m_i-y_i-1}\,d\,\theta_i \\ &= \frac{m_i!}{y_i!\,(m_i-y_i)!}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\frac{\Gamma(\alpha+y_i)\,\Gamma(\beta+m_i-y_i)}{\Gamma(\alpha+\beta+m_i)} \qquad (15.7) \end{aligned}$$

The joint mixture for $Y_1,\ldots,Y_n$ is the product of these pmfs,

$$h(\boldsymbol{y}|\alpha,\beta) = \prod_{i=1}^{n} h(y_i|\alpha,\beta). \qquad (15.8)$$

It is important here to keep track of the sets of possible values for all of the various quantities involved in these derivations. We have,

1. In $f(y_i|\theta_i)$, $y_i \in \Omega_Y = \{0,1,\ldots,m_i\}$ and $\theta_i \in \Theta = (0,\,1)$.
2. In $g(\theta_i|\alpha,\beta)$, $\theta_i \in \Theta = (0,\,1)$ and $\alpha > 0$, $\beta > 0$.
3. In $h(y_i|\alpha,\beta)$, $y_i \in \Omega_Y = \{0,1,\ldots,m_i\}$ and $\alpha > 0$, $\beta > 0$.

It is crucial that these sets of possible values (for y_i, θ_i, α, β) all match throughout the progression. Thus, the function $h(\cdot)$ is a probability mass function for the discrete random variable Y_i, and the derivation of $h(y_i|\alpha,\beta)$ has not changed the set of possible values from $f(y_i|\theta_i)$. If it had, our model would not make sense.

Now, in any estimation method we use, the log likelihood formed from the probability mass functions in (15.7) will be important (e.g., method of moments, maximum likelihood or Bayesian estimation). Using independence (Y_is conditionally independent given the θ_is, and the θ_is *iid* implies that marginally the Y_is are *iid*) we have that the log likelihood is,

$$\ell(\alpha,\beta) = \log\{h(\boldsymbol{y}|\alpha,\beta)\} = \sum_{i=1}^{n} \log\{h(y_i|\alpha,\beta)\}. \qquad (15.9)$$

The scientific mechanism or phenomenon of interest is the effect of Selenium on reproductive success in *Gambusia* and is embodied in the parameters α and β of the mixture model with log likelihood (15.9) which is written for one group. Our objective is to fit this model using both likelihood and Bayesian estimation.

Bayesian Analysis

In a Bayesian analysis we need to assign a joint prior to the fixed parameters α and β. The parameter space is $(\alpha, \beta) \in (0, \infty) \times (0, \infty)$ and it is difficult to determine an appropriate prior since we have no actual prior information available. In this case, a parameter transformation can make it easier to determine a naive prior. Let

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \phi = \frac{1}{\alpha + \beta + 1}. \tag{15.10}$$

The parameter space for (μ, ϕ) is now $(0, 1) \times (0, 1)$ and we might assign a joint prior to μ and ϕ as the product of two uniform distributions on the unit interval. Expressing the marginal probability mass functions (15.7) or (15.16) in terms of these parameters does nothing to simplify those expressions, but in writing computer functions we can simply use (15.9) after defining

$$\begin{aligned} \alpha &= \frac{(1-\phi)\mu}{\phi} \\ \beta &= \frac{(1-\phi)(1-\mu)}{\phi} \\ \alpha + \beta &= \frac{1-\phi}{\phi}. \end{aligned} \tag{15.11}$$

The joint posterior is then

$$p(\mu, \phi | \boldsymbol{y}) \propto h(\boldsymbol{y} | \mu, \phi) I(0 < \mu < 1)\, I(0 < \phi < 1), \tag{15.12}$$

where $I(A)$ is the indicator functions that assumes a value of 1 if A is true and a value of 0 otherwise. This joint posterior will need to be assessed through simulation rather than analytical derivation. We could certainly consider a Gibbs Sampling algorithm here, considering (15.12) first as a function of μ for a given ϕ and then again as a function of ϕ for a given μ. This would, however, perhaps be more cumbersome than needed, and we might like to simulate from a joint distribution with density proportional to (15.12) directly. An approach by which to accomplish this is provided by a Metropolis-Hastings algorithm.

To simulate from $p(\mu,\, \phi|\boldsymbol{y})$ using a Metropolis-Hastings algorithm, we need (1) a candidate distribution from which to produce proposed "jumps" for the sampler and (2) calculation of the probability for accepting proposed jumps. The first of these is fairly easy in this problem because the joint sample space of $(\mu,\, \phi)$ is $(0,\, 1) \times (0,\, 1)$, which suggests an independence chain with candidate distribution

$$f(\mu,\, \phi) = \begin{cases} 1 & \text{if } (\mu,\, \phi) \in\ (0,\, 1) \times (0,\, 1) \\ 0 & \qquad\qquad o.w. \end{cases} \tag{15.13}$$

We may easily simulate values $(\mu^*,\, \phi^*)$ from this distribution by simulating μ^* and ϕ^* independently from uniform distributions on the unit interval.

The Metropolis acceptance probability for proposed jumps from a current value $(\mu,\, \phi)$ to a new proposed value $(\mu^*,\, \phi^*)$ takes the form of

$$\begin{aligned} \alpha'[(\mu,\, \phi),\, (\mu^*,\, \phi^*)] &= \min\{h(\boldsymbol{y}|\mu^*,\, \phi^*)/h(\boldsymbol{y}|\mu,\, \phi),\ 1\} \\ &= \min\{w(\mu^*,\, \phi^*,\, \mu,\, \phi),\ 1\}, \end{aligned} \tag{15.14}$$

where we have denoted this probability as α' so as not to confuse it with the parameter α in the mixture model.

The specific form of (15.14) can be determined by noting that the candidate distribution may be considered to define an original Metropolis-Hastings algorithm in which the candidate distributions from the numerator and denominator of the acceptance probability cancel. Our essential difficulty at this point is computation of the ratio $w(\mu^*, \phi^*, \mu, \phi)$, which is complicated by the fact that $h(\boldsymbol{y}|\mu, \phi)$ in (15.7)) contains ratios of products of gamma functions. Such functions can easily assume either huge or negligible values, resulting in computational values of infinity or values that fail to exist (i.e., the `NaN` assignment in `R`). As a result, even though the ratios may be well within normal computational range, the components in the numerator or denominator may not be, producing computation algorithms that fail.

Our solution to this difficulty rests on two computational techniques that are both worth knowing. First, note that the form of $w(\cdot, \cdot)$ is a ratio, and any ratio may be written as the exponentiation of the difference of logarithms. Specifically,

$$\begin{aligned} w(\mu^*, \phi^*, \mu, \phi) &= \frac{h(\boldsymbol{y}|\mu^*, \phi^*)}{h(\boldsymbol{y}|\mu, \phi)} \\ &= \exp\left[\log\{h(\boldsymbol{y}|\mu^*, \phi^*)\} - \log\{h(\boldsymbol{y}|\mu, \phi)\}\right]. \quad (15.15) \end{aligned}$$

and $\log\{h(\boldsymbol{y}|\mu, \phi)\}$ can be computed as in (15.9) after applying the transformations (15.11). The second computational device comes from noticing that the components of (15.9) can be simplified by applying the property of gamma functions that $\Gamma(x) = (x-1)\Gamma(x-1)$ to the component densities in (15.7), which can then be written as

$$h(y_i|\alpha,\beta) = \begin{cases} \dfrac{\displaystyle\prod_{j=0}^{m_i-1}(\beta+j)}{\displaystyle\prod_{j=0}^{m_i-1}(\alpha+\beta+j)} & y_i = 0 \\ \frac{m_i!}{y_i!\,(m_i-y_i)!}\dfrac{\displaystyle\prod_{j=0}^{y_i-1}(\alpha+j)\prod_{j=0}^{m_i-y_i-1}(\beta+j)}{\displaystyle\prod_{j=0}^{m_i-1}(\alpha+\beta+j)} & 0 < y_i < m_i \\ \dfrac{\displaystyle\prod_{j=0}^{m_i-1}(\alpha+j)}{\displaystyle\prod_{j=0}^{m_i-1}(\alpha+\beta+j)} & y_i = n_i \end{cases} \qquad (15.16)$$

Notice that each product in (15.16) becomes a summation upon taking the logarithm.

All of this leads to a practical computational approach for implementation of a Metropolis algorithm to (1) generate candidate jumps and (2) calculate acceptance probabilities for those proposals. An outline of that algorithm is as follows.

1. Begin with initial values $(\mu_0,\, \phi_0) \in (0,\, 1) \times (0,\, 1)$

2. Set current values $(\mu_c,\, \phi_c) = (\mu_0,\, \phi_0)$ and set $t = 1$

3. At iteration t, generate a proposed jump $(\mu^*,\, \phi^*)$ as a pair of independent values from uniform distributions on the unit interval.

4. Compute $w(\mu^*,\, \phi^*,\, \mu_c,\, \phi_c)$ as given in (15.15) making use of (15.16) and (15.9) with $(\alpha_c,\, \beta_c)$ and $(\alpha^*,\, \beta^*)$ defined by the transformations in (15.11).

5. Generate an independent value u from a uniform distribution on (0, 1).

6. If $u \leq w(\mu^*, \phi^*, \mu_c, \phi_c)$ let $(\mu_t, \phi_t) = (\mu^*, \phi^*)$. Otherwise, let $(\mu_t, \phi_t) = (\mu_c, \phi_c)$.

7. Set $(\mu_c, \phi_c) = (\mu_t, \phi_t)$ and update t to $t + 1$, and return to step 3.

8. Discard values for a burn-in period $(t \leq B)$

9. Continue for M additional iterations, collecting values of (μ_t, ϕ_t) at each iteration.

At the conclusion of this algorithm we have a collection of M values of (μ, ϕ) simulated from the posterior $p(\mu, \phi|\boldsymbol{y})$.

This algorithm was applied to the *Gambusia* data from the San Luis Drain (SLD) and Volta areas using a burn-in of $B = 50$ and a total of $M = 50,000$ kept values. Starting values were $\mu_0 = 0.5$ and $\phi = 0.5$. Histograms of the posterior distributions of μ are presented in Figure 15.1 and those for ϕ are presented in Figure 15.2. Monte Carlo approximations to the posterior means, variances, and 90% credible intervals are contained in Table 15.1.

Area	Parameter	Mean	Variance	90% Interval
SLD	μ	0.799	0.0029	(0.702, 0.878)
Volta	μ	0.863	0.0039	(0.747, 0.944)
SLD	ϕ	0.245	0.0078	(0.123, 0.403)
Volta	ϕ	0.593	0.0188	(0.353, 0.804)

Table 15.1: Monte Carlo approximations to summary values from posterior distributions for μ and ϕ from the *Gambusia* reproduction study.

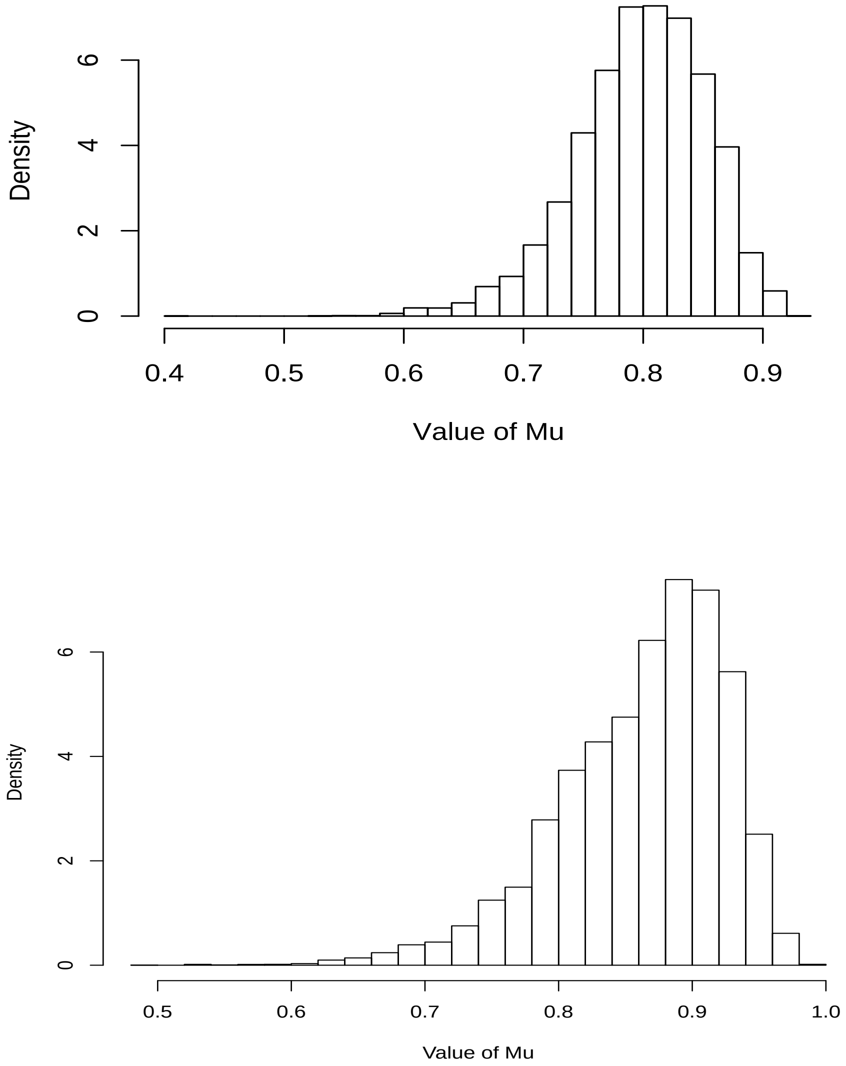

Figure 15.1: Posterior distribution of μ from the SLD (upper) and Volta (lower) areas in California.

Figure 15.1 and Table 15.1 indicate that the posterior distributions for μ are quite similar between the two areas. Figure 15.2 and Table 15.1 indicate

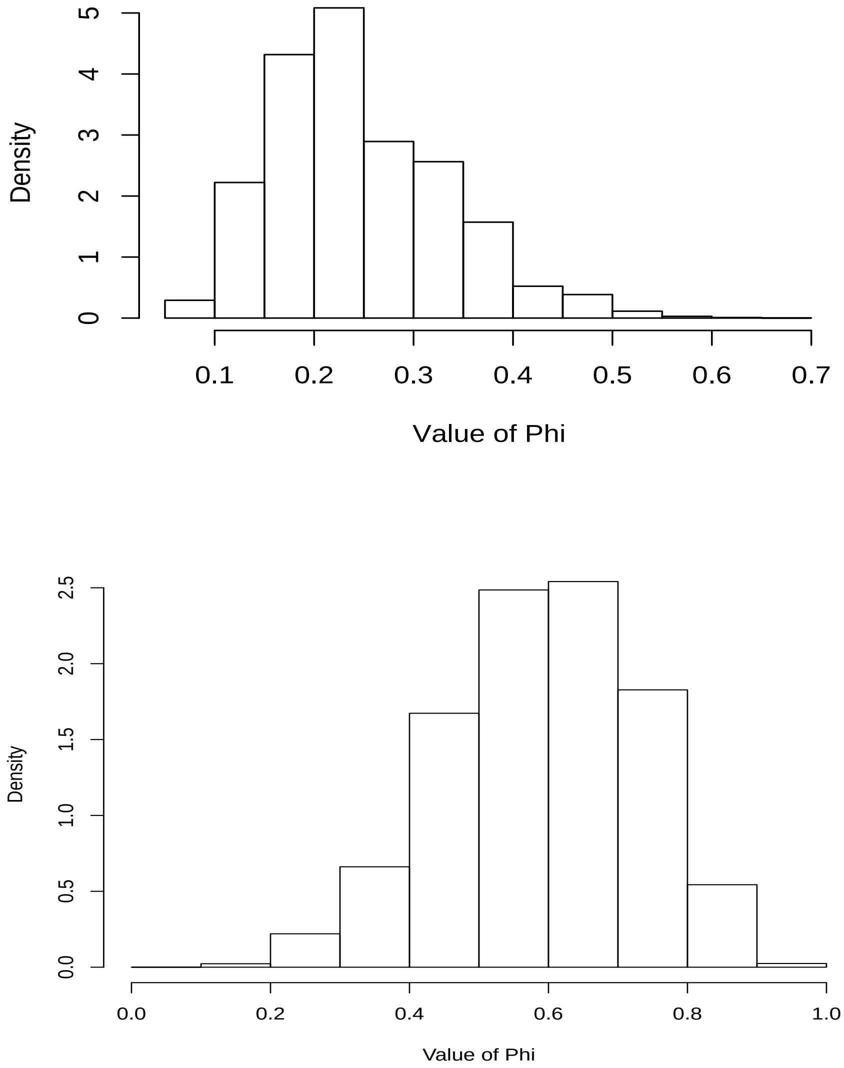

Figure 15.2: Posterior distribution of ϕ from the SLD (upper) and Volta (lower) areas in California.

that there seems to be some difference between the areas in terms of the parameter ϕ, although 90% credible intervals for this parameter do overlap.

Recall that the mechanism of interest is captured in the model in the beta mixing distribution (15.6), that is, the distribution of the $\{\theta_i;\ i = 1, \ldots, m\}$. The summary values of Table 15.1 demonstrate that there does not seem to be a difference between the SLD and Volta areas in terms of what we know about the location of these distributions (μ). And, it is difficult to interpret the seeming difference in values of ϕ directly, as this parameter does not describe a single characteristic of a beta distribution.

The solution from a Bayesian perspective is to examine the posterior predictive distribution of the probabilities of live births for the two areas, $p(\theta^*|\boldsymbol{y})$. This distribution is determined as

$$p(\theta^*|\boldsymbol{y}) = \int g(\theta^*|\mu, \phi)\, p(\mu, \phi|\boldsymbol{y})\, d\mu\ d\,\phi. \tag{15.17}$$

The expression (15.17) results from the fact that, by model assumption the distribution of θ given μ and ϕ is the same as the distribution of θ given μ, ϕ, and $\boldsymbol{y}$. The Metropolis-Hastings algorithm described previously produces values of (μ, ϕ) from the posterior $p(\mu,\ \phi|\boldsymbol{y})$. Based on (15.17) we may then simulate from the posterior predictive of the θ_i using the following algorithm.

1. For each pair of values (μ_m, ϕ_m) simulated from the posterior of these quantities, transform to (α_m, β_m) using the relations in expression (15.11).

2. Simulate a value θ^*_m set from $g(\theta|\alpha_m,\ \beta_m)$ which can be accomplished with the built-in R function for generating observations from a beta distribution.

3. Conducted for all $m = 1, \ldots, M$ pairs of values retained in th Metropolis-Hastings algorithm this results in a set of M values $\{\theta^*_m;\ m = 1, \ldots, M\}$ from the posterior predictive distribution (15.17).

A total of 50, 000 data sets were simulated from the posterior predictive distributions of both the SLD and Volta areas. Empirical distribution functions of these simulated values are presented in Figure 15.3.

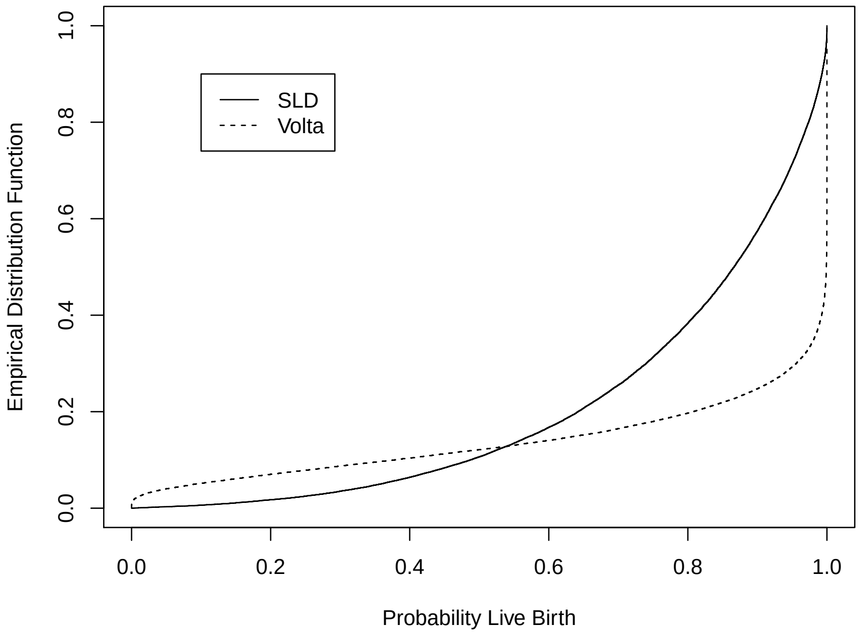

Figure 15.3: Empirical distribution functions of 50, 000 values simulated from the posterior predictive distribution for the SLD (solid line) and Volta (dashed line) areas.

Figure 15.3 and the numerical values it is drawn from indicate the difference in tendencies for live birth between the SLD and Volta areas. Based on these distributions we would conclude that the probability a fish from the SLD area has a probability of live birth less than 0.75 is 0.31 and the probability of live birth less than 0.90 is 0.57. In contrast, for the Volta area these probabilities are only 0.18 and 0.25, respectively. Thus, the overall conclusion would be that fish from the SLD area with high levels of Se are more likely than fish from the Volta area to exhibit poor reproductive success.

Likelihood Analysis

Direct maximization of the log likelihood (15.9) results in maximum likelihood estimates of μ and ϕ or, by invariance, α and β. Computations can again be made more stable by making use of (15.16) and computer functions can make use of either parameterization through the relations in (15.11). A 90% Wald theory interval for ϕ from the Volta area extended onto the negative line, suggesting the use of normed profile for computation of confidence intervals in this problem. A summary of maximum likelihood estimation is given in Table 15.2. These estimates compare favorably with the Bayesian posterior means, variances, and 90% credible intervals from Table 15.1. The greatest difference appears to be in estimation of ϕ from the Volta area and, in particular, the estimated variance of the approximate sampling distribution in Table 15.2 compared to the variance of the posterior distribution in Table 15.1. Of course, there is no reason these values should necessarily be in agreement. But it is comforting to find that two disparate but reasonable approaches to fitting the

Area	Parameter	Mean	Variance	90% Interval
SLD	μ	0.822	0.0023	(0.744, 0.901)
Volta	μ	0.889	0.0035	(0.792, 0.986)
SLD	ϕ	0.244	0.0140	(0.049, 0.439)
Volta	ϕ	1.341	0.9087	(0.297, 0.814)

Table 15.2: A summary of maximum likelihood estimation for the *Gambusia* reproduction study.

model lead to about the same results.

The maximized log likelihood for the SLD area was $\ell_s(\hat{\mu}, \hat{\phi}) = -136.8254$ and for the Volta area was $\ell_v(\hat{\mu}, \hat{\phi}) = -53.4091$. For all of the data combined the maximized log likelihood was $\ell(\hat{\mu}, \hat{\phi}) = -195.9984$. These values lead to the likelihood ratio test statistic

$$\begin{aligned} T &= -2\{-195.9984 - (-53.4091 + -136.8254)\} \\ &= 11.5277, \end{aligned}$$

which has an associated $p-$value of 0.00314 when compared to a Chi-squared distribution with 2 degrees of freedom. We would conclude that the full model with two beta mixing distributions is to be preferred to the reduced model with a single beta mixing distribution. A plot of the estimated mixing distributions for the two areas is presented in Figure 15.4 which is strikingly similar to the posterior predictive distributions in Figure 15.3, although not exactly the same.

15.5 Two Views of Hierarchical Models

Most statisticians, including the author of this book, would characterize the Bayesian analysis of a beta-binomial model in the *Gambusia* case study as an application of hierarchical modeling. There are several viewpoints under which hierarchical models can be conceptualized. These viewpoints do not affect the mathematics of analysis, but they do affect the process of making inference and, in particular, what we believe has been learned about a problem based on a statistical analysis. Before we present two of these viewpoints we first outline an alternative procedure by which we could simulate from the exact same posterior distributions as graphed in Figures 15.1 and 15.2.

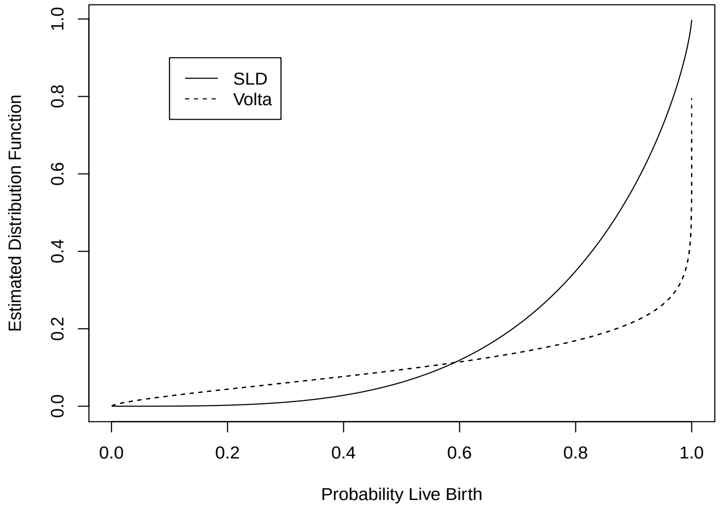

Figure 15.4: Estimated beta mixing distributions for the SLD area (solid curve) and Volta area (dashed curve) in California.

15.5.1 Integration from Simulation

In our analysis of the *Gambusia* case study we simulated the joint posterior through the use of a Metropolis-Hastings algorithm after having collapsed the original binomial data model and the beta mixing distribution to determine the marginal log likelihood analytically. It would also have been possible to make use of an MCMC algorithm without having first evaluated these integrals on paper. This could be done by a massive Metropolis-Hastings algorithm but is probably more easily seen by considering an overall Gibbs Sampling structure. Instead of forming the joint marginal distribution of responses as in (15.8) we will form the joint data model distribution

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{m} f_i(y_i|\theta_i),$$

where the $f_i(y_i|\theta_i)$ are binomial probability mass functions as in (15.5). Similarly, the joint mixing distribution is

$$g(\boldsymbol{\theta}|\alpha,\beta) = \prod_{i=1}^{m} g(\theta_i|\alpha,\beta),$$

where the $g(\theta_i|\alpha,\beta)$ are beta probability density functions as in (15.6). We will again consider this distribution to be parameterized by μ and ϕ from (15.10) and write $g(\boldsymbol{\theta}|\mu,\phi)$. Combining these two distributions with the joint prior $\pi(\mu,\phi)$ then gives a posterior as

$$p(\mu,\phi,\theta_1,\ldots,\theta_m|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\mu,\phi)\, \pi(\mu,\phi). \qquad (15.18)$$

Notice here that all of the data model parameters $\theta_1,\ldots,\theta_m$ are included in this posterior distribution. To implement a Gibbs algorithm for sampling from (15.18)we will need to determine full conditional distributions for each of the arguments on the left hand side of (15.18). This is accomplished by taking each argument one at a time and examining the right hand side of (15.18) to see what portions contain the relevant argument. We will write $p(x|\cdot)$ to denote the density of X conditioned on all other quantities that may be present in a problem. The full conditional densities needed here are then

$$\begin{aligned} p(\mu|\cdot) &\propto \pi(\mu,\phi)g(\boldsymbol{\theta}|\mu,\phi) \\ p(\phi|\cdot) &\propto \pi(\mu,\phi)g(\boldsymbol{\theta}|\mu,\phi) \\ \text{and, for } i = 1,\ldots,n & \\ p(\theta_i|\cdot) &\propto g(\theta_i|\mu,\phi)\, f_i(y_i|\theta_i) \qquad (15.19) \end{aligned}$$

Several points should be made about the densities in (15.19). First, notice that there are $n + 2$ distributions represented, but they are all univariate. Notice that the distributions on the right hand side in $p(\theta_i|\cdot)$ are fairly simple, involving only one θ_i and one y_i. Finally, be aware that although the right hand side of the expressions for $p(\mu|\cdot)$ and $p(\phi|\cdot)$ are the same these are two different distributions. A Gibbs Sampling algorithm for this problem can be outlined as follows.

1. Choose starting values μ_0, ϕ_0, $\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{n,0}$, and values for burn-in B and eventual sample size M. Set $t = 1$.

2. At iteration t,

   (a) Simulate a value μ_t from $p(\mu|\phi_{t-1}, \theta_{1,t-1}, \ldots, \theta_{n,t-1})$

   (b) Simulate a value ϕ_t from $p(\phi|\mu_t, \theta_{1,t-1}, \ldots, \theta_{n,t-1})$

   (c) For $i = 1, \ldots, n$ simulate a value of θ_i from $p(\theta_i|\mu_t, \phi_t, y_i)$

   Update $t = t + 1$ and repeat for $t = 1, \ldots, B + M$. Discard values from iterations $t <= B$.

At termination of this algorithm we will have M values of the vector

$$(\mu, \phi, \theta_1, \ldots, \theta_n)$$

simulated from its joint posterior distribution (15.18). Recall that simulation from joint distributions also accomplishes simulation from marginal distributions. Thus, if we take the M values of μ or ϕ we should get histograms essentially the same as those in Figures 15.1 and 15.2 and summary values as in Table 15.1.

Notice, however, that we also have M values simulated from the distributions $p(\theta_1|\boldsymbol{y})$, $p(\theta_2|\boldsymbol{y}), \ldots, p(\theta_n|\boldsymbol{y})$. A question is whether these distributions are of use in making inferences about the problem, and it is opinion about this question that distinguishes the two viewpoints of hierarchical models that are the topic of this section.

15.5.2 Hierarchical Models as Mixtures

We have given a quite general formulation of general mixture models. One view is that a hierarchical model really is nothing more than a mixture model with priors assigned to the parameters of the mixing distribution. There are some situations for which this view is difficult to maintain, primarily in problems that involve the use of dynamic models that evolve over time. But for many, many problems this way of thinking about hierarchical models seems to have much to recommend it. In this context, a hierarchical model consists of three pieces,

$$\begin{array}{ll} \text{Data Model:} & f(\boldsymbol{y}|\boldsymbol{\theta}) \\ \text{Mixing Distribution:} & g(\boldsymbol{\theta}|\boldsymbol{\lambda}) \\ \text{Prior:} & \pi(\boldsymbol{\lambda}). \end{array}$$

Although we may use a Gibbs Sampling algorithm to simulate from the joint distribution $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y})$, interest is typically only in $p(\boldsymbol{\lambda}|\boldsymbol{y})$ for the purposes of inference. The values in $\boldsymbol{\theta}$ are considered as random variables that play the role of parameters in the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$, but we are not interested in what values these quantities had. The focus of scientific inference is often the distribution of these variables, in that it governs the way a scientific mechanism of interest manifests itself, as discussed more fully in Chapter 15.1. As a result, we are also generally interested in the posterior predictive distribution $p(\boldsymbol{\theta}^*|\boldsymbol{y})$, as given for the *Gambusia* problem in (15.17).

Although we may not always and, in fact, seldom do derive a marginal data model analytically for a Bayesian analysis, we can consider this view of hierarchical models as producing a data model that is a general mixture model and then conducting a Bayesian analysis in the usual way by assigning a prior distribution to the fixed parameter. This then reduces the three pieces listed previous to only two pieces,

$$\begin{array}{ll} \text{Data Model:} & f(\boldsymbol{y}|\boldsymbol{\lambda}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta} \\ \text{Prior:} & \pi(\boldsymbol{\lambda}). \end{array}$$

15.5.3 Hierarchical Models as Models with Multi-stage Priors

Another view of hierarchical models is that such models are formulated for sets of similar problems, each of which might have the same form of data model but with different parameter values, $f_i(y_i|\boldsymbol{\theta}_i)$. We might like to assign each of these data models the same prior $\pi(\boldsymbol{\theta}_i)$ but are uncertain what values to pick for any parameters that may be present in the prior selected. Thus, we allow the parameters of this *first-stage* prior to remain unspecified, and assign them an additional completely specified *second-stage* prior distribution. In this context, a hierarchical model consists of three pieces

$$\begin{array}{ll} \text{Data Model:} & f(\boldsymbol{y}|\boldsymbol{\theta}) \\ \text{Stage 1 Prior:} & \pi_1(\boldsymbol{\theta}|\boldsymbol{\lambda}) \\ \text{Stage 2 Prior:} & \pi_2(\boldsymbol{\lambda}). \end{array}$$

Notice that with $\pi_1 = g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ the three pieces are identical to those listed for the viewpoint in which hierarchical models are a Bayesian version of mixture models, they are just considered to have different roles. Here, we still have the joint posterior proportional to the product of the three pieces, we are still able to simulate from $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y})$ and examine any margin that we wish. In other words, nothing has changed mathematically. What has changed is that often we are interested in posterior distributions of the individual data model parameters $p(\boldsymbol{\theta}_i|\boldsymbol{y})$ or components of $\boldsymbol{\theta}_i$.

Here, the problems of interest are contained in the data model and the hierarchical structure can be thought of as a method for arriving at a common prior for those problems. The three pieces of a hierarchical model again reduce to two pieces, but this time as

$$\begin{array}{ll} \text{Data Model:} & f(\boldsymbol{y}|\boldsymbol{\theta}) \\ \text{Prior:} & \pi(\boldsymbol{\theta}) = \int \pi_1(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi_2(\boldsymbol{\lambda})\, d\,\boldsymbol{\lambda}, \end{array}$$

and this should be contrasted with the Data Model and Prior of the previous subsection.

15.5.4 Resolving the Two Viewpoints

Some statisticians claim that hierarchical models are justified on the basis of what are called *representation theorems*, which essentially indicate that any reasonable distribution can be represented as a mixture model. Representation theorems are interesting mathematical results (that can actually become quite involved), but they do not resolve questions about how a hierarchical model is conceptualizing a problem. These questions are inherently philosophical and scientific in nature, not mathematical. If one believes that mathematics can provide guidance on this issue, then one must also be willing to construct mathematical arguments for or against the use of other distributions that can be derived from a hierarchical model, such as the conditional posterior $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\lambda})$, which has an interesting history in the Bayesian literature and was seemingly more widely accepted as a reasonable distribution to examine for inference before the advent of MCMC methods than it is now. At the same time, it is, in the opinion of the author, not productive to argue about which of the viewpoints about hierarchical models presented is “correct”, “better”, or “more justified”. Rather, the question should be one of what aspects of a problem indicate that one or the other viewpoint is to be preferred for that particular problem.

All settings in which hierarchical models have application involve observations in more than one setting or situation. It is these situations to which one value of the data model parameters are applicable. If one observed only a single situation then one would have a traditional data model with a single fixed parameter value. If one has observed all of the situations of interest, then one has a collection of problems, each of which would lend itself to inference through use of the posterior distribution of its parameter values. This would fall under the viewpoint of hierarchical models as models with multi-stage priors. If one has a random sample of the situations of interest, one is likely more interested in the distribution of data model parameters (or random variables that play the role of parameters in the data model) than in the particular values pertaining to each observed situation. This would lend itself to inference through the view of hierarchical models as arising from the Bayesian analysis of general mixture models. Unfortunately, it will be rare that one every has observed every situation of interest or that one has a true random sample from some large population of situations. Thus, while these pure scenarios are valuable in considering distinctions between viewpoints of hierarchical models they cannot provide a guideline for what to do in practice.

A meaningful consideration when contemplating whether individual data model parameters are to be a subject of inference is whether or not it would be at least hypothetically possible to obtain another observation from the data models for particular situations, that is, the data model with the same values of the parameters that led to the observed data. This will perhaps become more clear by considering a hypothetical scenario involving a type of cardiac therapy intended to prevent second heart attacks. Suppose observations on the number of patients having and not having second heart attacks is available over a set of medical facilities that offer this therapy. A hierarchical model very similar to that used in the analysis of *Gambusia* reproductive success would be appropriate for this problem. Here, the θ_i are probabilities of a second heart attack for patients being treated at facility $i = 1, \ldots, n$. The θ_i are assigned a beta distribution with parameters μ and ϕ and the joint posterior $p(\mu, \phi, \theta_1, \ldots, \theta_n|\boldsymbol{y})$ is determined through simulation. Now, the question to be addressed concerns which posterior distributions should be used to make inference. The marginal posterior $p(\mu|\boldsymbol{y})$ is almost certainly of interest as it concerns the manner in which the mechanism of interest, which is the efficacy of the therapy in preventing heart attacks, manifests itself in different medical facilities of the types observed. If would certainly be beneficial for this inference if the facilities observed could be considered representative of some larger population. More problematic are the posteriors $\{p(\theta_i|\boldsymbol{y});\ \ i = 1, \ldots, m\}$. The argument in favor of using these distributions for inference is that a hypothetical patient who has a choice of which medical facility at which to enroll in the therapy might like to know how those facilities rank in effectiveness. A caveat that points directly back to the first sentence of this paragraph is that this would only be true if conditions, staff and equipment at the facilities were the same at the time that patient makes his or her choice as when the original data were collected. That is, if the patient believes he or she will become another observation from the data models with the same parameters that were in force at the time of the original analysis. One only need change heart attacks at medical facilities to success in placement of graduates for graduate programs, or performance of stocks over the past quarter, or any number of other scenarios to see the generality of this prescription for determining which posterior distributions are appropriate for inference in a problem.