distributed model (5.1), so that the conditional posterior density is a product of conjugate posterior densities for the components θ_j.

The third step can be performed by brute force by integrating the joint posterior distribution over θ:

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta. \tag{5.4}$$

For many standard models, however, including the normal distribution, the marginal posterior distribution of ϕ can be computed algebraically using the conditional probability formula,

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}. \tag{5.5}$$

This expression is useful because the numerator is just the joint posterior distribution (5.3), and the denominator is the posterior distribution for θ if ϕ were known. The difficulty in using (5.5), beyond a few standard conjugate models, is that the denominator, $p(\theta|\phi, y)$, regarded as a function of both θ and ϕ for fixed y, has a normalizing factor that depends on ϕ as well as y. One must be careful with the proportionality 'constant' in Bayes' theorem, especially when using hierarchical models, to make sure it is actually constant. Exercise 5.11 has an example of a nonconjugate model in which the integral (5.4) has no closed-form solution so that (5.5) is no help.

Drawing simulations from the posterior distribution

The following strategy is useful for simulating a draw from the joint posterior distribution, $p(\theta, \phi|y)$, for simple hierarchical models such as are considered in this chapter.

1. Draw the vector of hyperparameters, ϕ, from its marginal posterior distribution, $p(\phi|y)$. If ϕ is low-dimensional, the methods discussed in Chapter 3 can be used; for high-dimensional ϕ, more sophisticated methods such as described in Part III may be needed.
2. Draw the parameter vector θ from its conditional posterior distribution, $p(\theta|\phi, y)$, given the drawn value of ϕ. For the examples we consider in this chapter, the factorization $p(\theta|\phi, y) = \prod_j p(\theta_j|\phi, y)$ holds, and so the components θ_j can be drawn independently, one at a time.
3. If desired, draw predictive values $\tilde{y}$ from the posterior predictive distribution given the drawn θ. Depending on the problem, it might be necessary first to draw a new value $\tilde{\theta}$, given ϕ, as discussed at the end of the previous section.

As usual, the above steps are performed L times in order to obtain a set of L draws. From the joint posterior simulations of θ and $\tilde{y}$, we can compute the posterior distribution of any estimand or predictive quantity of interest.

Application to the model for rat tumors

We now perform a full Bayesian analysis of the rat tumor experiments described in Section 5.1. Once again, the data from experiments $j = 1, \ldots, J$, $J = 71$, are assumed to follow independent binomial distributions:

$$y_j \sim \text{Bin}(n_j, \theta_j),$$

with the number of rats, n_j, known. The parameters θ_j are assumed to be independent samples from a beta distribution:

$$\theta_j \sim \text{Beta}(\alpha, \beta),$$

and we shall assign a noninformative hyperprior distribution to reflect our ignorance about the unknown hyperparameters. As usual, the word 'noninformative' indicates our attitude toward this part of the model and is not intended to imply that this particular distribution has any special properties. If the hyperprior distribution turns out to be crucial for our inference, we should report this and if possible seek further substantive knowledge that could be used to construct a more informative prior distribution. If we wish to assign an improper prior distribution for the hyperparameters, (α, β), we must check that the posterior distribution is proper. We defer the choice of noninformative hyperprior distribution, a relatively arbitrary and unimportant part of this particular analysis, until we inspect the integrability of the posterior density.

Joint, conditional, and marginal posterior distributions. We first perform the three steps for determining the analytic form of the posterior distribution. The joint posterior distribution of all parameters is

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta) \\ &\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}. \end{aligned} \tag{5.6}$$

Given (α, β), the components of θ have independent posterior densities that are of the form $\theta_j^A(1-\theta_j)^B$—that is, beta densities—and the joint density is

$$p(\theta|\alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}. \tag{5.7}$$

We can determine the marginal posterior distribution of (α, β) by substituting (5.6) and (5.7) into the conditional probability formula (5.5):

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}. \tag{5.8}$$

The product in equation (5.8) cannot be simplified analytically but is easy to compute for any specified values of (α, β) using a standard routine to compute the gamma function.

Choosing a standard parameterization and setting up a 'noninformative' hyperprior distribution. Because we have no immediately available information about the distribution of tumor rates in populations of rats, we seek a relatively diffuse hyperprior distribution for (α, β). Before assigning a hyperprior distribution, we reparameterize in terms of $\text{logit}(\frac{\alpha}{\alpha+\beta}) = \log(\frac{\alpha}{\beta})$ and $\log(\alpha+\beta)$, which are the logit of the mean and the logarithm of the 'sample size' in the beta population distribution for θ. It would seem reasonable to assign independent hyperprior distributions to the prior mean and 'sample size,' and we use the logistic and logarithmic transformations to put each on a $(-\infty, \infty)$ scale. Unfortunately, a uniform prior density on these newly transformed parameters yields an improper *posterior* density, with an infinite integral in the limit $(\alpha+\beta) \to \infty$, and so this particular prior density cannot be used here.

In a problem such as this with a reasonably large amount of data, it is possible to set up a 'noninformative' hyperprior density that is dominated by the likelihood and yields a proper posterior distribution. One reasonable choice of diffuse hyperprior density is uniform on $(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2})$, which when multiplied by the appropriate Jacobian yields the following densities on the original scale,

$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}, \tag{5.9}$$

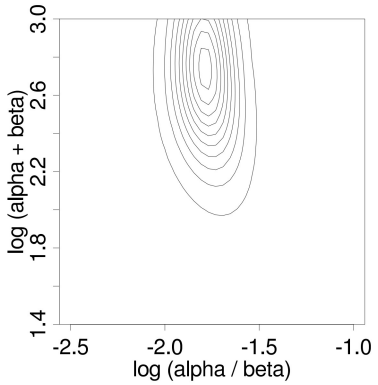

Figure 5.2 *First try at a contour plot of the marginal posterior density of* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *for the rat tumor example. Contour lines are at* $0.05, 0.15, \ldots, 0.95$ *times the density at the mode.*

and on the natural transformed scale:

$$p\left(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2}. \tag{5.10}$$

See Exercise 5.9 for a discussion of this prior density.

We could avoid the mathematical effort of checking the integrability of the posterior density if we were to use a proper hyperprior distribution. Another approach would be tentatively to use a flat hyperprior density, such as $p(\frac{\alpha}{\alpha+\beta}, \alpha+\beta) \propto 1$, or even $p(\alpha, \beta) \propto 1$, and then compute the contours and simulations from the posterior density (as detailed below). The result would clearly show the posterior contours drifting off toward infinity, indicating that the posterior density is not integrable in that limit. The prior distribution would then have to be altered to obtain an integrable posterior density.

Incidentally, setting the prior distribution for $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ to uniform in a vague but finite range, such as $[-10^{10}, 10^{10}] \times [-10^{10}, 10^{10}]$, would *not* be an acceptable solution for this problem, as almost all the posterior mass in this case would be in the range of α and β near 'infinity,' which corresponds to a Beta(α, β) distribution with a variance of zero, meaning that all the θ_j parameters would be essentially equal in the posterior distribution. When the likelihood is not integrable, setting a faraway finite cutoff to a uniform prior density does not necessarily eliminate the problem.

Computing the marginal posterior density of the hyperparameters. Now that we have established a full probability model for data and parameters, we compute the marginal posterior distribution of the hyperparameters. Figure 5.2 shows a contour plot of the unnormalized marginal posterior density on a grid of values of $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$. To create the plot, we first compute the logarithm of the density function (5.8) with prior density (5.9), multiplying by the Jacobian to obtain the density $p(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)|y)$. We set a grid in the range $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)) \in [-2.5, -1] \times [1.5, 3]$, which is centered near our earlier point estimate $(-1.8, 2.3)$ (that is, $(\alpha, \beta) = (1.4, 8.6)$) and covers a factor of 4 in each parameter. Then, to avoid computational overflows, we subtract the maximum value of the log density from each point on the grid and exponentiate, yielding values of the unnormalized marginal posterior density.

The most obvious features of the contour plot are (1) the mode is not far from the point estimate (as we would expect), and (2) important parts of the marginal posterior distribution lie outside the range of the graph.

We recompute $p(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)|y)$, this time in the range $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)) \in$

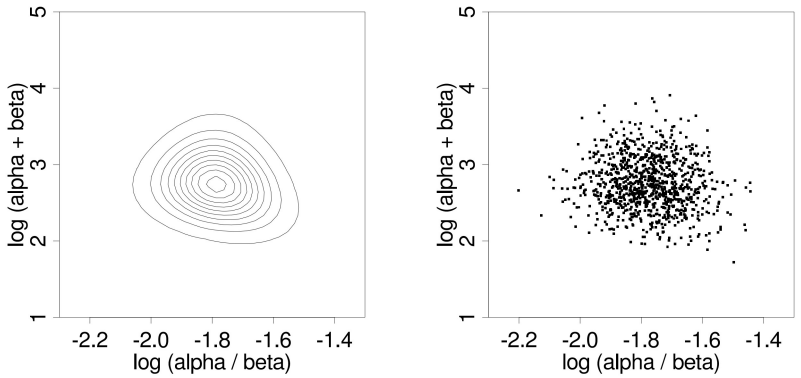

Figure 5.3 *(a) Contour plot of the marginal posterior density of* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *for the rat tumor example. Contour lines are at* $0.05, 0.15, \ldots, 0.95$ *times the density at the mode. (b) Scatterplot of 1000 draws* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *from the numerically computed marginal posterior density.*

$[-2.3, -1.3] \times [1, 5]$. The resulting grid, shown in Figure 5.3a, displays essentially all of the marginal posterior distribution. Figure 5.3b displays 1000 random draws from the numerically computed posterior distribution. The graphs show that the marginal posterior distribution of the hyperparameters, under this transformation, is approximately symmetric about the mode, roughly $(-1.75, 2.8)$. This corresponds to approximate values of $(\alpha, \beta) = (2.4, 14.0)$, which differs somewhat from the crude estimate obtained earlier.

Having computed the relative posterior density at a grid that covers the effective range of (α, β), we normalize by approximating the distribution as a step function over the grid and setting the total probability in the grid to 1.

We can then compute posterior moments based on the grid of $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$; for example,

$$\mathrm{E}(\alpha|y) \text{ is estimated by } \sum_{\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)} \alpha \cdot p(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)|y).$$

From the grid in Figure 5.3, we compute $\mathrm{E}(\alpha|y) = 2.4$ and $\mathrm{E}(\beta|y) = 14.3$. This is close to the estimate based on the mode of Figure 5.3a, given above, because the posterior distribution is approximately symmetric on the scale of $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$. A more important consequence of averaging over the grid is to account for the posterior uncertainty in (α, β), which is not captured in the point estimate.

Sampling from the joint posterior distribution of parameters and hyperparameters. We draw 1000 random samples from the joint posterior distribution of $(\alpha, \beta, \theta_1, \ldots, \theta_J)$, as follows.

1. Simulate 1000 draws of $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ from their posterior distribution displayed in Figure 5.3, using the same discrete-grid sampling procedure used to draw (α, β) for Figure 3.3b in the bioassay example of Section 3.8.
2. For $l = 1, \ldots, 1000$:
   (a) Transform the lth draw of $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ to the scale (α, β) to yield a draw of the hyperparameters from their marginal posterior distribution.
   (b) For each $j = 1, \ldots, J$, sample θ_j from its conditional posterior distribution, $\theta_j|\alpha, \beta, y \sim \mathrm{Beta}(\alpha + y_j, \beta + n_j - y_j)$.

Displaying the results. Figure 5.4 shows posterior medians and 95% intervals for the θ_j's, computed by simulation. The rates θ_j are shrunk from their sample point estimates, $\frac{y_j}{n_j}$,

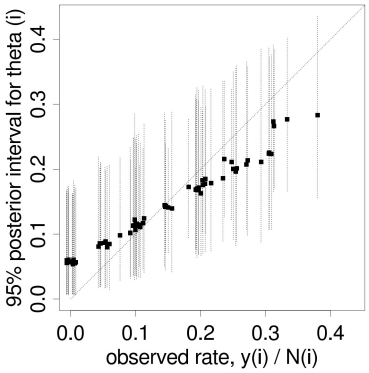

Figure 5.4 *Posterior medians and 95% intervals of rat tumor rates, θ_j (plotted vs. observed tumor rates y_j/n_j), based on simulations from the joint posterior distribution. The 45° line corresponds to the unpooled estimates, $\hat{\theta}_i = y_i/n_i$. The horizontal positions of the line have been jittered to reduce overlap.*

towards the population distribution, with approximate mean 0.14; experiments with fewer observations are shrunk more and have higher posterior variances. The results are superficially similar to what would be obtained based on a point estimate of the hyperparameters, which makes sense in this example, because of the fairly large number of experiments. But key differences remain, notably that posterior variability is higher in the full Bayesian analysis, reflecting posterior uncertainty in the hyperparameters.

5.4 Normal model with exchangeable parameters

We now present a full treatment of a simple hierarchical model based on the normal distribution, in which observed data are normally distributed with a different mean for each 'group' or 'experiment,' with known observation variance, and a normal population distribution for the group means. This model is sometimes termed the one-way normal random-effects model with known data variance and is widely applicable, being an important special case of the hierarchical normal linear model, which we treat in some generality in Chapter 15. In this section, we present a general treatment following the computational approach of Section 5.3. The following section presents a detailed example; those impatient with the algebraic details may wish to look ahead at the example for motivation.

The data structure

Consider J independent experiments, with experiment j estimating the parameter θ_j from n_j independent normally distributed data points, y_{ij}, each with known error variance σ^2; that is,

$$y_{ij}|\theta_j \sim \mathrm{N}(\theta_j, \sigma^2), \text{ for } i = 1, \ldots, n_j; \;\; j = 1, \ldots, J. \tag{5.11}$$

Using standard notation from the analysis of variance, we label the sample mean of each group j as

$$\overline{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

with sampling variance

$$\sigma_j^2 = \sigma^2/n_j.$$