Example. Estimating the risk of tumor in a group of rats
In the evaluation of drugs for possible clinical application, studies are routinely performed on rodents. For a particular study drawn from the statistical literature, suppose the immediate aim is to estimate θ, the probability of tumor in a population of female laboratory rats of type 'F344' that receive a zero dose of the drug (a control group). The data show that 4 out of 14 rats developed endometrial stromal polyps (a kind of tumor). It is natural to assume a binomial model for the number of tumors, given θ. For convenience, we select a prior distribution for θ from the conjugate family, $\theta \sim \text{Beta}(\alpha, \beta)$.

Analysis with a fixed prior distribution. From historical data, suppose we knew that the tumor probabilities θ among groups of female lab rats of type F344 follow an approximate beta distribution, with known mean and standard deviation. The tumor probabilities θ vary because of differences in rats and experimental conditions among the experiments. Referring to the expressions for the mean and variance of the beta distribution (see Appendix A), we could find values for α, β that correspond to the given values for the mean and standard deviation. Then, assuming a $\text{Beta}(\alpha, \beta)$ prior distribution for θ yields a $\text{Beta}(\alpha + 4, \beta + 10)$ posterior distribution for θ.

Approximate estimate of the population distribution using the historical data. Typically, the mean and standard deviation of underlying tumor risks are not available. Rather, historical *data* are available on previous experiments on similar groups of rats. In the rat tumor example, the historical data were in fact a set of observations of tumor incidence in 70 groups of rats (Table 5.1). In the jth historical experiment, let the number of rats with tumors be y_j and the total number of rats be n_j. We model the y_j's as independent binomial data, given sample sizes n_j and study-specific means θ_j. Assuming that the beta prior distribution with parameters (α, β) is a good description of the population distribution of the θ_j's in the historical experiments, we can display the hierarchical model schematically as in Figure 5.1, with θ_{71} and y_{71} corresponding to the current experiment.

The observed sample mean and standard deviation of the 70 values y_j/n_j are 0.136 and 0.103. If we set the mean and standard deviation of the population distribution to these values, we can solve for α and β—see (A.3) on page 582 in Appendix A. The resulting estimate for (α, β) is $(1.4, 8.6)$. This is *not* a Bayesian calculation because it is not based on any specified full probability model. We present a better, fully Bayesian approach to estimating (α, β) for this example in Section 5.3. The estimate $(1.4, 8.6)$ is simply a starting point from which we can explore the idea of estimating the parameters of the population distribution.

Using the simple estimate of the historical population distribution as a prior distribution for the current experiment yields a $\text{Beta}(5.4, 18.6)$ posterior distribution for θ_{71}: the posterior mean is 0.223, and the standard deviation is 0.083. The prior

Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:

4/14

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of y_j/n_j: (number of rats with tumors)/(total number of rats).*

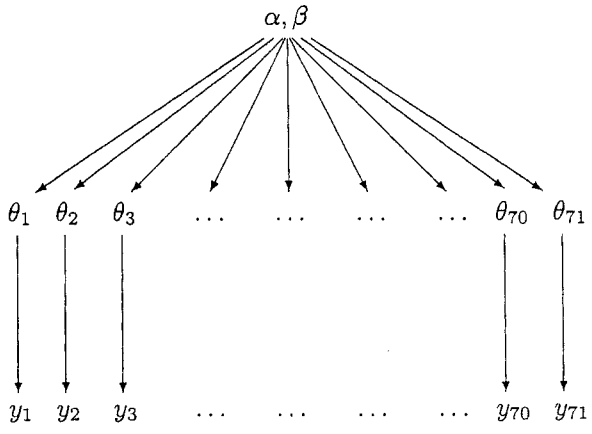

Figure 5.1 *Structure of the hierarchical model for the rat tumor example.*

information has resulted in a posterior mean substantially lower than the crude proportion, $4/14 = 0.286$, because the weight of experience indicates that the number of tumors in the current experiment is unusually high.

These analyses require that the current tumor risk, θ_{71}, and the 70 historical tumor risks, $\theta_1, \ldots, \theta_{70}$, be considered a random sample from a common distribution, an assumption that would be invalidated, for example, if it were known that the historical experiments were all done in laboratory A but the current data were gathered in laboratory B, or if time trends were relevant. In practice, a simple, although arbitrary, way of accounting for differences between the current and historical data is to inflate the historical variance. For the beta model, inflating the historical variance means decreasing $(\alpha+\beta)$ while holding α/β constant. Other systematic differences, such as a time trend in tumor risks, can be incorporated in a more extensive model.

Having used the 70 historical experiments to form a prior distribution for θ_{71}, we might now like also to use this same prior distribution to obtain Bayesian inferences for the tumor probabilities in the first 70 experiments, $\theta_1, \ldots, \theta_{70}$. There are several logical and practical problems with the approach of directly estimating a prior distribution from existing data:

- If we wanted to use the estimated prior distribution for inference about the first 70 experiments, then the data would be used twice: first, all the results together are used to estimate the prior distribution, and then each experiment's results are used to estimate its θ. This would seem to cause us to overestimate our precision.
- The point estimate for α and β seems arbitrary, and using any point estimate for α and β necessarily ignores some posterior uncertainty.
- We can also make the opposite point: does it make sense to 'estimate' α and β at all? They are part of the 'prior' distribution: should they be known before the data are gathered, according to the logic of Bayesian inference?

Logic of combining information

Despite these problems, it clearly makes more sense to try to estimate the population distribution from all the data, and thereby to help estimate each θ_j, than to estimate all 71 values θ_j separately. Consider the following thought experiment about inference on two of the parameters, θ_{26} and θ_{27}, each corresponding to experiments with 2 observed tumors out of 20 rats. Suppose our prior distribution for both θ_{26} and θ_{27} is centered around 0.15; now suppose that you were told after completing the data analysis that $\theta_{26} = 0.1$ exactly. This should influence your estimate of θ_{27}; in fact, it would probably make you think that θ_{27} is lower than you previously believed, since the data for the two parameters are identical, and the postulated value of 0.1 is lower than you previously expected for θ_{26} from the prior distribution. Thus, θ_{26} and θ_{27} should be dependent in the posterior distribution, and they should not be analyzed separately.

We retain the advantages of using the data to estimate prior parameters and eliminate all of the disadvantages just mentioned by putting a probability model on the entire set of parameters and experiments and then performing a Bayesian analysis on the joint distribution of all the model parameters. A complete Bayesian analysis is described in Section 5.3. The analysis using the data to estimate the prior parameters, which is sometimes called *empirical Bayes*, can be viewed as an approximation to the complete hierarchical Bayesian analysis. We prefer to avoid the term 'empirical Bayes' because it misleadingly suggests that the full Bayesian method, which we discuss here and use for the rest of the book, is not 'empirical.'

Example. Rat tumors (continued)
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

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

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

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

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

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

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,

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

and on the natural transformed scale:

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

See Exercise 5.7 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

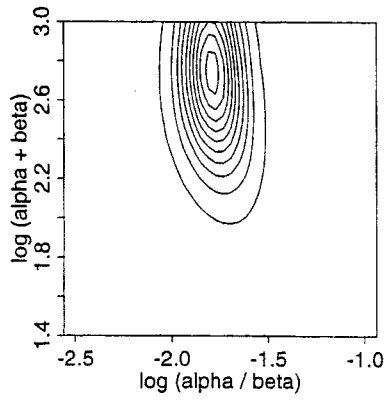

Figure 5.2 *First try at a contour plot of the marginal posterior density of* $(\log(\frac{\alpha}{\beta}), \log(\alpha + \beta))$ *for the rat tumor example. Contour lines are at* $0.05, 0.15, \ldots, 0.95$ *times the density at the mode.*

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 the Jacobian to obtain the density $p(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)|y)$. We set a grid in the range $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)) \in [-1, -2.5] \times [1.5, 3]$, 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 $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)) \in [-1.3, -2.3] \times [1, 5]$. 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.

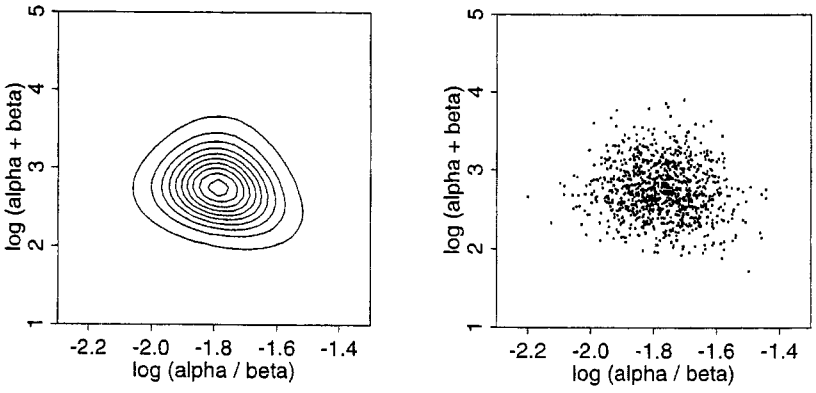

Figure 5.3 *(a) Contour plot of the marginal posterior density of* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *for the rat tumor example. Contour lines are at* $0.05, 0.15, \ldots, 0.95$ *times the density at the mode. (b) Scatterplot of 1000 draws* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$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

Having computed the relative posterior density at a grid of values that cover 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,

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

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 sample (α, β) for Figure 3.4b 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, $\theta_j|\alpha, \beta, y \sim \mathrm{Beta}(\alpha + y_j, \beta + n_j - y_j)$.

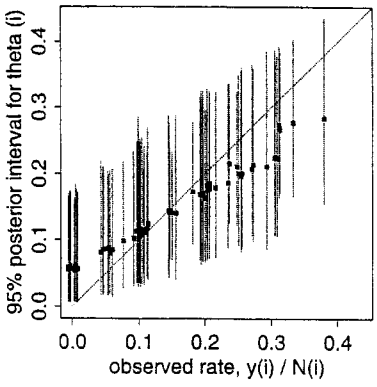

Figure 5.4 *Posterior medians and 95% intervals of rat tumor rates, θ_j (plotted vs. observed tumor rates y_j/n_j), based on simulations from the joint posterior distribution. The 45° line corresponds to the unpooled estimates, $\hat{\theta}_i = y_i/n_i$. The horizontal positions of the line have been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

Displaying the results. Figure 5.4 shows posterior means and 95% intervals for the θ_j's, computed by simulation. The rates θ_j are shrunk from their sample point estimates, y_j/n_j, 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.