7

Hierarchical Modeling

7.1 Introduction

In this chapter, we illustrate the use of R to summarize an exchangeable hierarchical model. We begin by giving a brief introduction to hierarchical modeling. Then we consider the simultaneous estimation of the true mortality rates from heart transplants for a large number of hospitals. Some of the individual estimated mortality rates are based on limited data and it may be desirable to combine the individual rates in some way to obtain more accurate estimates. We describe a two-stage model, a mixture of gamma distributions, to represent prior beliefs that the true mortality rates are exchangeable. We describe the use of R to simulate from the posterior distribution. We first use contour graphs and simulation to learn about the posterior distribution of the hyperparameters. Once we simulate hyperparameters, we can simulate from the posterior distributions of the true mortality rates from gamma distributions. We conclude by illustrating how the simulation of the joint posterior can be used to perform different types of inferences in the heart transplant application.

7.2 Introduction to Hierarchical Modeling

In many statistical problems, we are interested in learning about many parameters that are connected in some way. To illustrate, consider the following three problems described in this chapter and the chapters to follow.

1. **Simultaneous estimation of hospital mortality rates**
   In the main example of this chapter, one is interested in learning about the mortality rates due to heart transplant surgery for 94 hospitals. Each hospital has a true mortality rate λ_i, and so one wishes to simultaneously estimate the 94 rates $\lambda_1, ..., \lambda_{94}$. It is reasonable to believe a priori that the true rates are similar in size, which implies a dependence structure

between the parameters. If one is told some information about a particular hospital's true rate, that information would likely affect one's belief about the location of a second hospital's rate.

2. **Estimating college grade point averages**
   In an example in Chapter 10, admissions people at a particular university collect a table of means of freshman grade point averages (GPA) organized by the student's high school rank and his or her score on a standardized test. One wishes to learn about the collection of population mean GPAs with the ultimate goal of making predictions about the success of future students that attend the university. One believes that the population GPAs can be represented as a simple linear function of the high school rank and standardized test score.
3. **Estimating career trajectories**
   In an example in Chapter 11, one is learning about the pattern of performance of athletes as they age during their sports careers. In particular, one wishes to estimate the *career trajectories* of the batting performances of a number of baseball players. For each player, one fits a model to estimate his career trajectory, and Fig. 7.1 displays the fitted career trajectories for nine players. Note that the shapes of these trajectories are similar; a player generally will increase in performance until his late 20s or early 30s and then decline until retirement. The prior belief is that the true trajectories will be similar between players, which again implies a prior distribution with dependence.

In many-parameter situations like the ones described here, it is natural to construct a prior distribution in a *hierarchical* fashion. In this type of model, the observations are given distributions conditional on parameters, and the parameters in turn have distributions conditional on additional parameters called hyperparameters. Specifically, we begin by specifying a data distribution

$$y \sim f(y|\theta),$$

and the prior vector θ will be assigned a prior distribution with unknown hyperparameters λ:

$$\theta \sim g_1(\theta|\lambda).$$

The hyperparameter vector λ in turn will be assigned a distribution

$$\lambda \sim g_2(\lambda).$$

One general way of constructing a hierarchical prior is based on the prior belief of *exchangeability*. A set of parameters $\theta = (\theta_1, ..., \theta_k)$ is exchangeable if the distribution of θ is unchanged if the parameter components are permuted. This implies that one's prior belief about θ_j, say, will be the same as one's belief about θ_h. One can construct an exchangeable prior by assuming that the components of θ are a random sample from a distribution g_1:

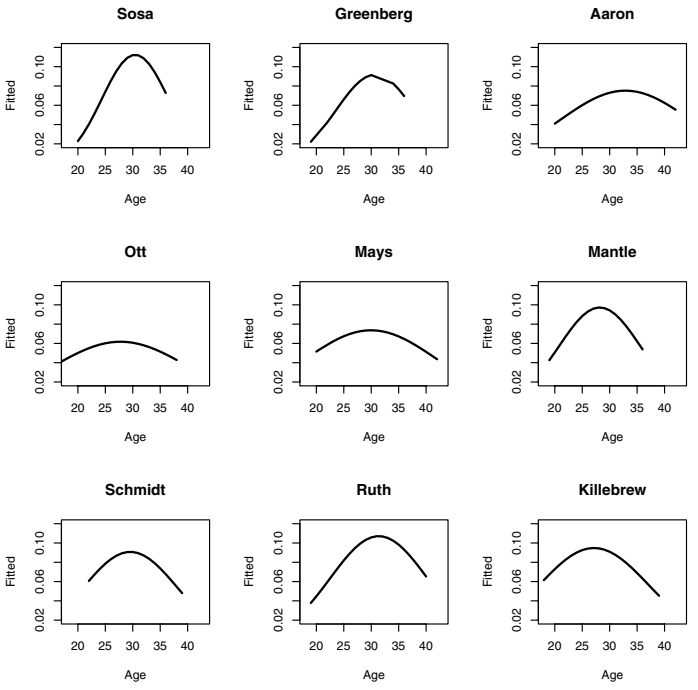

Fig. 7.1. Plots of fitted career trajectories for nine baseball players as a function of their age.

$$\theta_1, ..., \theta_k \text{ random sample from } g_1(\theta|\lambda),$$

and the unknown hyperparameter vector λ is assigned a known prior at the second stage:

$$\lambda \sim g_2(\lambda).$$

This particular form of hierarchical prior will be used for the mortality rates example of this chapter and the career trajectories example of Chapter 11.

7.3 Individual and Combined Estimates

Consider again the heart transplant mortality data discussed in Chapter 3. The number of deaths within 30 days of heart transplant surgery is recorded for each of 94 hospitals. In addition, we record for each hospital an expected number of deaths called the exposure denoted by e. We let y_i and e_i denote the respective observed number of deaths and exposure for the ith hospital. In R, we read in the relevant dataset `hearttransplants` in the LearnBayes package.

```
> data(hearttransplants)
> attach(hearttransplants)
```

A standard model assumes that the number of deaths y_i follows a Poisson distribution with mean $e_i\lambda_i$ and the objective is to estimate the mortality rate per unit exposure λ_i. The fraction y_i/e_i is the number of deaths per unit exposure and can be viewed as an estimate of the death rate for the ith hospital. In Fig. 7.2, we plot the ratios $\{y_i/e_i\}$ against the logarithms of the exposures $\{\log(e_i)\}$ for all hospitals where each point is labeled by the number of observed deaths y_i.

```
> plot(log(e), y/e, pch = as.character(y))
```

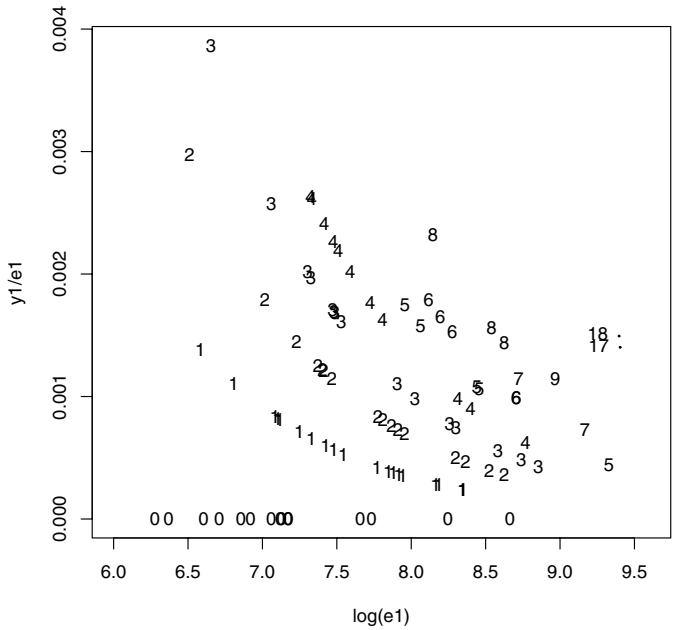

Fig. 7.2. Plot of death rates against log exposure for all hospitals. Each point is labeled by the number of observed deaths.

Note that the estimated rates are highly variable, especially for programs with small exposures. The programs experiencing no deaths (a plotting label of 0) also are primarily associated with small exposures.

Suppose we are interested in simultaneously estimating the true mortality rates $\{\lambda_i\}$ for all hospitals. One option is to simply estimate the true rates by

the individual death rates

$$\frac{y_1}{e_1}, ..., \frac{y_{94}}{e_{94}}.$$

Unfortunately these individual rates can be poor estimates, especially for the hospitals with small exposures. In Fig. 7.2, we saw that some of these hospitals did not experience any deaths and the individual death rate $y_i/e_i = 0$ would likely underestimate the hospital's true mortality rate. Also it is clear from the figure that the rates for the hospitals with small exposures have high variability.

Since the individual death rates can be poor, it seems desirable to combine the individual estimates in some way to obtain improved estimates. Suppose we can assume that the true mortality rates are equal across hospitals; that is,

$$\lambda_1 = ... = \lambda_{94}.$$

Under this "equal-means" Poisson model, the estimate of the mortality rate for the ith hospital would be the pooled estimate

$$\frac{\sum_{j=1}^{94} y_j}{\sum_{j=1}^{94} e_j}.$$

But this pooled estimate is based on the strong assumption that the true mortality rate is the same across hospitals. This is questionable since one would expect some variation in the true rates.

We have discussed two possible estimates for the mortality rate of the ith hospital: the individual estimate y_i/e_i and the pooled estimate $\sum y_j / \sum e_j$. A third possibility is the compromise estimate

$$(1 - \lambda)\frac{y_i}{e_i} + \lambda \frac{\sum_{j=1}^{94} y_j}{\sum_{j=1}^{94} e_j}.$$

This estimate shrinks or moves the individual estimate y_i/e_i toward the pooled estimate $\sum y_j / \sum e_j$ where the parameter $0 < \lambda < 1$ determines the size of the shrinkage. We will see that this shrinkage estimate is a natural byproduct of the application of an exchangeable prior model on the true mortality rates.

7.4 Equal Mortality Rates?

Before we consider an exchangeable model, let's illustrate fitting and checking the model where the mortality rates are assumed equal. Suppose y_i is distributed Poisson$(e_i\lambda)$, $i = 1, ..., 94$, and the common mortality rate λ is assigned a standard noninformative prior of the form

$$g(\lambda) \propto \frac{1}{\lambda}.$$

Then the posterior density of λ is given by

$$g(\lambda|\text{data}) \propto \frac{1}{\lambda} \prod_{j=1}^{94} \left[\lambda^{y_j} \exp(-e_j\lambda)\right]$$
$$= \lambda^{\sum_{j=1}^{94} y_j - 1} \exp(-\sum_{j=1}^{94} e_j\lambda)$$

which is recognized as a gamma density with parameters $\sum_{j=1}^{94} y_j$ and $\sum_{j=1}^{94} e_j$. For our data, we compute

```
> sum(y)

[1] 277

> sum(e)

[1] 294681
```

and so the posterior density for the common rate λ is gamma(277, 294681).

One general Bayesian method of checking the suitability of a fitted model such as this is based on the posterior predictive distribution. Let y_i^* denote the number of transplant deaths for hospital i with exposure e_i in a future sample. Conditional on the true rate λ, y_i^* has a Poisson distribution with mean $e_i\lambda$. Our current beliefs about the ith true rate are contained in the posterior density $g(\lambda|y)$. The unconditional distribution of y_i^*, the posterior predictive density, is given by

$$f(y_i^*|e_i, y) = \int f_P(y_i^*|e_i\lambda)g(\lambda|y)d\lambda,$$

where $f_P(y|\lambda)$ is the Poisson sampling density with mean λ. The posterior predictive density represents the likelihood of future observations based on our fitted model. For example, the density $f(y_i^*|e_i, y)$ represents the number of transplant deaths that we would predict in the future for a hospital with exposure e_i. If the actual number of observed deaths y_i is in the middle of this predictive distribution, then we can say that our observation is consistent with our model fit. On the other hand, if the observed y_i is in the extreme tails of the distribution $f(y_i^*|e_i, y)$, then this observation indicates that the model is inadequate in fitting this observation.

To illustrate the use of the posterior predictive distribution, consider hospital 94 that had 17 transplant deaths, that is, $y_{94} = 17$. Did this hospital have an unusually high number of deaths? To answer this question, we simulate 1000 values from the posterior predictive density of y_{94}^*.

To simulate from the predictive distribution of y_{94}^*, we first simulate 1000 draws of the posterior density of λ

```
> lambda=rgamma(1000,shape=277,rate=294681)
```

and then simulate draws of y_{94}^* from a Poisson distribution with mean $e_{94}\lambda$.

```
> ys94=rpois(1000,e[94]*lambda)
```

Using the following R code, Fig. 7.3 displays a histogram of this posterior predictive distribution and the actual number of transplant deaths y_{94} is shown by a vertical line.

```
> hist(ys94,breaks=seq(1.5,26.5,by=1))
> lines(c(y[94],y[94]),c(0,120),lwd=3)
```

Since the observed y_j is in the tail portion of the distribution, it seems inconsistent with the fitted model – it suggests that this hospital actually has a higher true mortality rate than estimated from this equal-rates model.

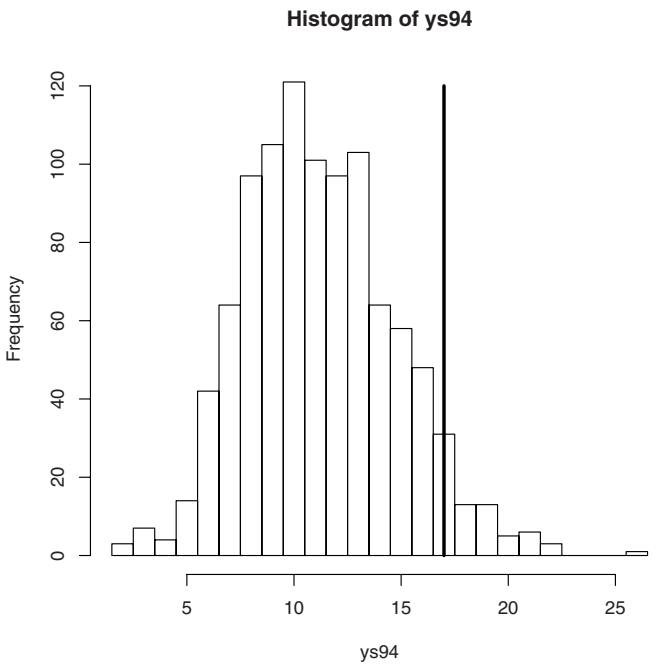

Fig. 7.3. Histogram of simulated draws from the posterior predictive distribution of y_{94}^*. The actual number of transplant deaths is shown by a vertical line.

We can check the consistency of the observed y_i with its posterior predictive distribution for all hospitals. For each distribution, we compute the probability that the future observation y_i^* is at least as extreme as y_i:

$$\min\{P(y_i^* \leq y_i), P(y_i^* \geq y_i)\}.$$

The following R code computes the probabilities of "at least as extreme" for all observations and places the probabilities in the vector `pout`.

```
> pout=0*y
> lambda=rgamma(1000,shape=277,rate=294681)
> for (i in 1:94){
+   ysi=rpois(1000,e[i]*lambda)
+   pleft=sum(ysi<=y[i])/1000
+   pright=sum(ysi>=y[i])/1000
+   pout[i]=min(pleft,pright)
+ }
```

We plot the probabilities against the log exposures which is displayed in Fig. 7.4.

```
> plot(log(e),pout,ylab="Prob(extreme)")
```

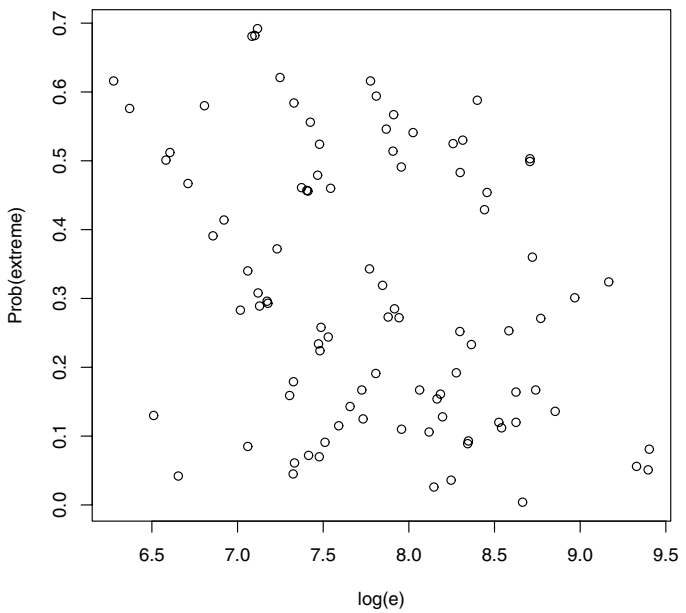

Fig. 7.4. Scatterplot of predictive probabilities of "at least as extreme" against log exposures for all observations.

Note that a number of these tail probabilities appear small (15 are smaller than 0.10) which means that the "equal rates" model is inadequate for explaining the distribution of mortality rates for the group of 94 hospitals. We will have to assume differences between the true mortality rates that will be modeled by the exchangeable model described in the next section.

7.5 Modeling a Prior Belief of Exchangeability

At the first stage of the prior, the true death rates $\lambda_1, ..., \lambda_{94}$ are assumed to be a random sample from a gamma$(\alpha, \alpha/\mu)$ distribution of the form

$$g(\lambda|\alpha, \mu) = \frac{(\alpha/\mu)^\alpha \lambda^{\alpha-1} \exp(-\alpha\lambda/\mu)}{\Gamma(\alpha)}, \lambda > 0.$$

The prior mean and variance of λ are given by μ and μ^2/α, respectively. At the second stage of the prior, the hyperparameters μ and α are assumed independent, with μ assigned a gamma(a, b) distribution with density $\mu^{a-1} \exp(-b\mu)$ and α the density $g(\alpha)$.

This prior distribution induces positive correlation between the true death rates. To see this, suppose one assigns the hyperparameter μ a gamma(10, 10) distribution and sets the hyperparameter α equal to a fixed value α_0. (This is equivalent to assigning a density $g(\alpha)$ that places probability one on the value α_0.) One can simulate values of, say (λ_1, λ_2), from the prior distribution by

- simulating values from μ from the gamma(a, b) distribution, α from the prior density $g(\alpha)$
- for each simulated pair (μ, α), simulate λ_1, λ_2 from gamma$(\alpha, \alpha/\mu)$ distributions

This simulation is illustrated in the following R code. Fig. 7.5 displays 500 simulated values from the prior distribution of (λ_1, λ_2) for the values α_0 equal to 5, 20, 80, and 400. Note that since μ is assigned a gamma(10, 10) distribution, both the true rates λ_1 and λ_2 are centered about the value 1. The hyperparameter α is a precision parameter that controls the correlation between the parameters. For the fixed value $\alpha = 400$, note that λ_1 and λ_2 are concentrated along the line $\lambda_1 = \lambda_2$. As the precision parameter α approaches infinity, the exchangeable prior places all of its mass along the space where $\lambda_1 = ... = \lambda_{94}$.

```
> par(mfrow = c(2, 2))
> m = 500
> alphas = c(5, 20, 80, 400)
> for (j in 1:4) {
+     mu = rgamma(m, shape = 10, rate = 10)
+     lambda1 = rgamma(m, shape=alphas[j], rate=alphas[j]/mu)
+     lambda2 = rgamma(m, shape=alphas[j], rate=alphas[j]/mu)
```

```
+     plot(lambda1, lambda2)
+     title(main=paste("alpha=",as.character(alphas[j])))
+ }
```

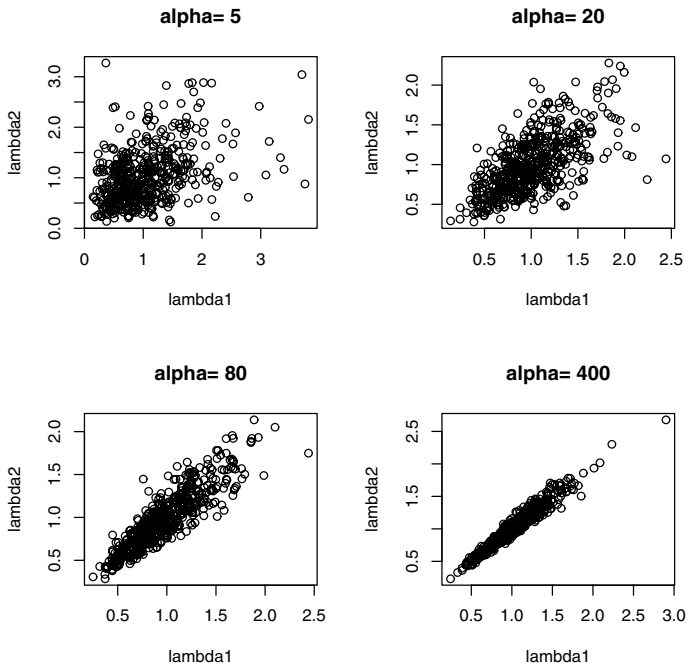

Fig. 7.5. Simulated values from the exchangeable prior on (λ_1, λ_2) for values of the precision parameter $\alpha = 5$, 20, 80, and 400.

Although we used subjective priors to illustrate the behavior of the prior distribution, in practice vague distributions can be chosen for the hyperparameters μ and α. In this example, we assign the mean parameter the typical vague prior of the form

$$g(\mu) \propto \frac{1}{\mu}, \mu > 0.$$

The precision parameter α assigned the proper, but relatively flat, prior density of the form

$$g(\alpha) = \frac{z_0}{(\alpha + z_0)^2}, \alpha > 0.$$

The user will specify a value of the parameter z_0 that is the median of α. In this example, we let $z_0 = 0.53$.

7.6 Posterior Distribution

Owing to the conditionally independent structure of the hierarchical model and the choice of a conjugate prior form at stage 2, there is a relatively simple posterior analysis. Conditional on values of the hyperparameters μ and α, the rates $\lambda_1, ..., \lambda_{94}$ have independent posterior distributions. The posterior distribution of λ_i is gamma$(y_i + \alpha, e_i + \alpha/\mu)$. The posterior mean of λ_i, conditional on α and μ, can be written as

$$E(\lambda_i|y, \alpha, \mu) = \frac{y_i + \alpha}{e_i + \alpha/\mu} = (1 - B_i)\frac{y_i}{e_i} + B_i\mu,$$

where

$$B_i = \frac{\alpha}{\alpha + e_i\mu}.$$

The posterior mean of the true rate λ_i can be viewed as a shrinkage estimator, where B_i is the shrinkage fraction of the posterior mean away from the usual estimate y_i/e_i toward the prior mean μ.

Also since a conjugate model structure was used, the rates λ_i can be integrated out of the joint posterior density, resulting in the marginal posterior density of (α, μ):

$$p(\alpha, \mu|\text{data}) = K\frac{1}{\Gamma^{94}(\alpha)}\prod_{j=1}^{94}\left[\frac{(\alpha/\mu)^\alpha\Gamma(\alpha + y_i)}{(\alpha/\mu + e_i)^{(\alpha+y_i)}}\right]\frac{z_0}{(\alpha + z_0)^2}\frac{1}{\mu},$$

where K is a proportionality constant.

7.7 Simulating from the Posterior

In the previous section the posterior density of all parameters was expressed as

$$g(\text{hyperparameters}|\text{data})\ g(\text{true rates}|\text{hyperparameters}, \text{data}),$$

where the hyperparameters are (μ, α) and the true rates are $(\lambda_1, ..., \lambda_{94})$. By the composition method, we can simulate a random draw from the joint posterior by

- simulating (μ, α) from the marginal posterior distribution
- simulating $\lambda_1, ..., \lambda_{94}$ from their distribution conditional on the values of the simulated μ and α

First we need to simulate from the marginal density of the hyperparameters μ and α. Since both parameters are positive, a good first step in this simulation process is to transform each to the real-valued parameters

$$\theta_1 = \log(\alpha), \theta_2 = \log(\mu).$$

The marginal posterior of the transformed parameters is given by

$$p(\theta_1, \theta_2|\text{data}) = K \frac{1}{\Gamma^{94}(\alpha)} \prod_{j=1}^{94} \left[\frac{(\alpha/\mu)^\alpha \Gamma(\alpha + y_i)}{(\alpha/\mu + e_i)^{(\alpha+y_i)}} \right] \frac{z_0 \alpha}{(\alpha + z_0)^2}.$$

The following R function `poissgamexch` contains the definition of the log posterior of θ_1 and θ_2.

```
poissgamexch=function(theta,datapar)
{
y=datapar$data[,2]; e=datapar$data[,1]
z0=datapar$z0
alpha=exp(theta[,1]); mu=exp(theta[,2])
beta=alpha/mu
N=length(y)
val=0*alpha;
for (i in 1:N)
{
val=val+lgamma(alpha+y[i])-(y[i]+alpha)*log(e[i]+beta)+
  alpha*log(beta)
}
val=val-N*lgamma(alpha)+log(alpha)-2*log(alpha+z0)
return(val)
}
```

Note that this function has two inputs:

- `theta`– a matrix of two columns where each row corresponds to a value of (θ_1, θ_2)
- `datapar` – a R list with two components, the `data` and the value of the hyperparameter `z0`

Note that since `theta` is a matrix, we sum over the observations to compute the log posterior. We use the function `lgamma` that computes the log of the gamma function, $\log \Gamma(x)$.

Using the R function `laplace`, we find the posterior mode and associated variance-covariance matrix. We perform five iterations of the Newton-Raphson algorithm at the starting value $(\theta_1, \theta_2) = (2, -7)$. The output of `laplace` includes the mode and the corresponding estimate at the variance-covariance matrix.

```
> datapar = list(data = hearttransplants, z0 = 0.53)
> start=array(c(2, -7), c(1, 2))
> fit = laplace(poissgamexch, start, 5, datapar)
> fit

$mode
        [,1]       [,2]
```

```
[1,] 1.88535 -6.955614

$var
            [,1]         [,2]
[1,]  0.23412668 -0.003077430
[2,] -0.00307743  0.005863179

$int
[1] -2208.502
```

This output gives us information about the location of the posterior density. By trial and error, we use the function `mycontour` to find a grid that contains the posterior density of (θ_1, θ_2). The resulting graph is displayed in Fig. 7.6.

```
> par(mfrow = c(1, 1))
> mycontour(poissgamexch, c(0, 8, -7.3, -6.6), datapar)
> title(xlab="log alpha",ylab="log mu")
```

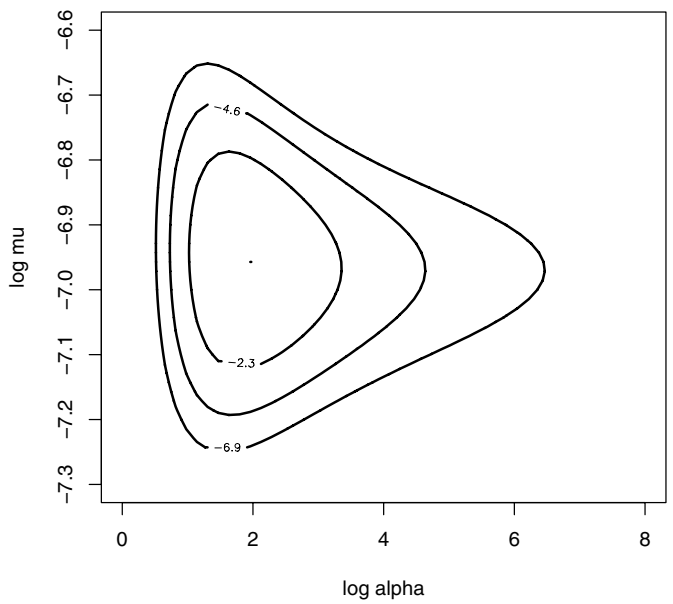

Fig. 7.6. Contour plot of the posterior density of $(\log \alpha, \log \mu)$ for the heart transplant example. Contour lines are drawn at 10%, 1%, and .1% of the modal value.

By inspection of Fig. 7.6, we see that the posterior density for (θ_1, θ_2) is nonnormal shaped, especially in the direction of $\theta_1 = \log \alpha$. Since the normal approximation to the posterior is inadequate, we obtain a simulated sample of (θ_1, θ_2) by use of the "Metropolis within Gibbs" algorithm in the function `gibbs`. In this Gibbs sampling algorithm, we start at the value $(\theta_1, \theta_2) = (4, -7)$ and iterate through 1000 cycles with Metropolis scale parameters $c_1 = 1, c_2 = .15$. As the output indicates, the acceptance rates in the simulation of the two conditional distributions are each about 30%.

```
> start = array(c(4, -7), c(1, 2))
> fitgibbs = gibbs(poissgamexch, start, 1000, c(1,.15), datapar)
> fitgibbs$accept

     [,1]  [,2]
[1,] 0.312 0.284
```

Fig. 7.7 shows a simulated sample of 1000 placed on top of the contour graph. Note that most of the points fall within the first two contour lines of the graph, indicating that the algorithm appears to give a representative sample from the marginal posterior distribution of θ_1 and θ_2.

```
> mycontour(poissgamexch, c(0, 8, -7.3, -6.6), datapar)
> points(fitgibbs$par[, 1], fitgibbs$par[, 2])
```

Fig. 7.8 shows a kernel density estimate of the simulated draws from the marginal posterior distribution of the precision parameter $\theta_1 = \log(\alpha)$.

```
> plot(density(fitgibbs$par[, 1], bw = 0.2))
```

We can learn about the true mortality rates $\lambda_1, ..., \lambda_{94}$ by simulating values of from their posterior distributions. Given values of the hyperparameters α and μ, the true rates have independent posterior distributions with λ_i distributed gamma$(y_i + \alpha, e_i + \alpha/\mu)$. For each rate, we use the R `rgamma` function to obtain a sample from the gamma distribution, where the gamma parameters are functions of the simulated values of α and μ. For example, one can obtain a sample from the posterior distribution of λ_1 by the R code

```
> alpha = exp(fitgibbs$par[, 1])
> mu = exp(fitgibbs$par[, 2])
> lam1 = rgamma(1000, y[1] + alpha, e[1] + alpha/mu)
```

After we obtain a simulated sample of size 1000 for each true rate λ_i, we can summarize each sample by computing the 5th and 95th percentiles. The interval from these two percentiles constitutes an approximate 90% probability interval for λ_i. We graph these 90% probability intervals as vertical lines on our original graph of the log exposures and the individual rates in Fig. 7.9. In contrast to the wide variation in the observed death rates, note the similarity in the locations of the probability intervals for the true rates. This indicates

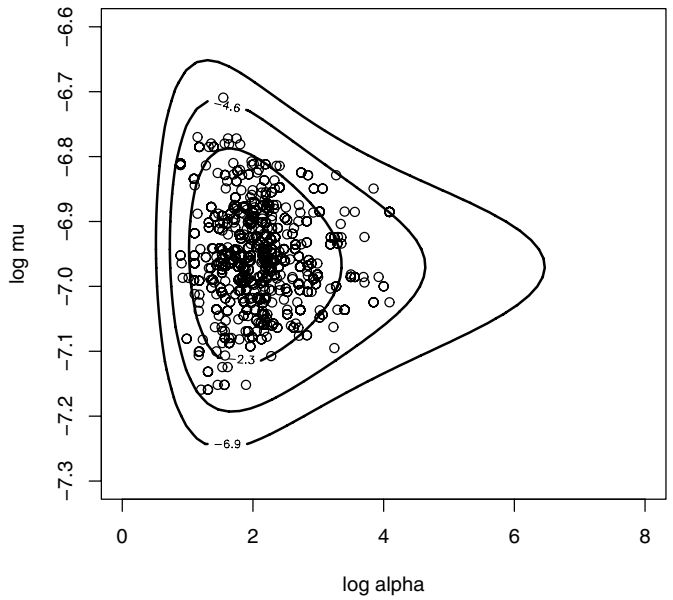

Fig. 7.7. Contour plot of the posterior density of $(\log \alpha, \log \mu)$ for the heart transplant example with a sample of simulated values placed on top.

that these Bayesian estimates are shrinking the individual rates toward the pooled estimate.

```
> alpha = exp(fitgibbs$par[, 1])
> mu = exp(fitgibbs$par[, 2])
> plot(log(e), y/e, pch = as.character(y))
> for (i in 1:94) {
+     lami = rgamma(1000, y[i] + alpha, e[i] + alpha/mu)
+     probint = quantile(lami, c(0.05, 0.95))
+     lines(log(e[i]) * c(1, 1), probint)
+ }
```

7.8 Posterior Inferences

Once a simulated sample of true rates $\{\lambda_i\}$ and the hyperparameters μ, α has been generated from the joint posterior distribution, we can use this sample to perform various types of inferences.

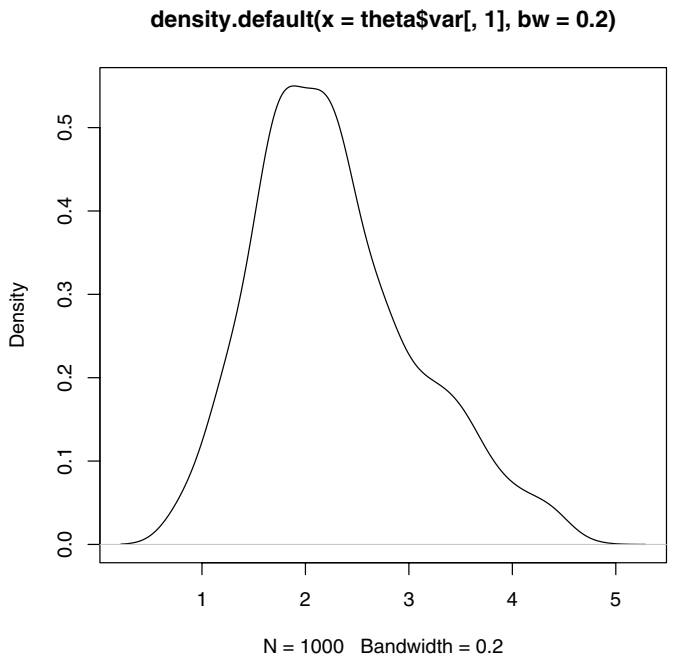

Fig. 7.8. Density estimate of simulated draws from the marginal posterior of $\log \alpha$.

7.8.1 Shrinkage

The posterior mean of the ith true mortality rate λ_i can be approximated by

$$E(\lambda_i|\text{data}) \approx (1 - E(B_i|\text{data}))\frac{y_i}{e_i} + E(B_i|\text{data})\frac{\sum_{j=1}^{94} y_j}{\sum_{j=1}^{94} e_j},$$

where $B_i = \alpha/(\alpha + e_i\mu)$ is the size of the shrinkage of the ith observed rate y_i/e_i toward the pooled estimate $\sum_{j=1}^{94} y_j / \sum_{j=1}^{94} e_j$. In the following R code, we compute the posterior mean of the shrinkage sizes $\{B_i\}$ for all 94 components. In Fig. 7.10, we plot the mean shrinkages against the logarithms of the exposures. For the hospitals with small exposures, the Bayesian estimate shrinks the individual estimate 90% toward the combined estimate. In contrast, for large hospitals with high exposures, the shrinkage size is closer to 50%.

```
> shrinkage = 0 * e
> for (i in 1:94) shrinkage[i] = mean(alpha/(alpha + e[i] * mu))
> plot(log(e), shrinkage)
```

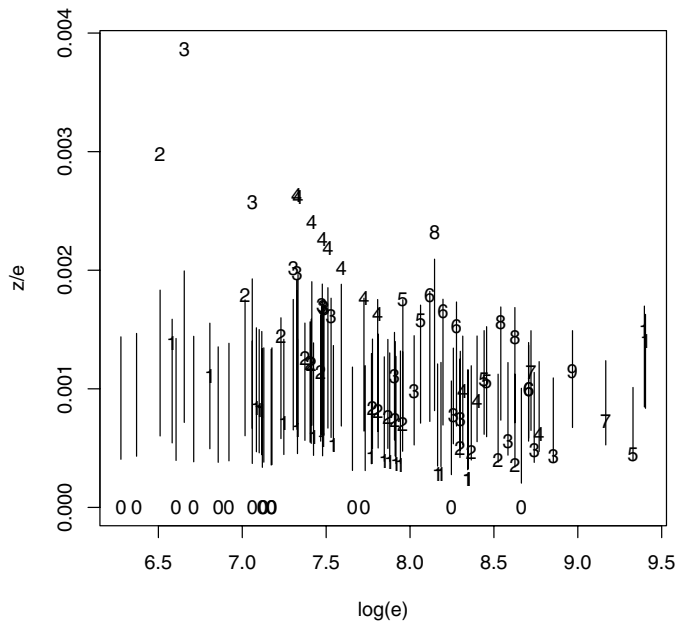

Fig. 7.9. Plot of observed death rates against log exposures together with intervals representing 90% posterior probability bands for the true rates $\{\lambda_i\}$.

7.8.2 Comparing Hospitals

Suppose one is interested in comparing the true mortality rates of the hospitals. Specifically, suppose one wishes to compare the "best hospital" with the other hospitals. First, we find the hospital with the smallest estimated mortality rate. In the following R output, we compute the posterior mean of the mortality rates, where the posterior mean of the true rate for hospital i is given by

$$E\left(\frac{y_i + \alpha}{e_i + \alpha/\mu}\right),$$

where the expectation is taken over the marginal posterior distribution of (α, μ).

```
> hospital=1:94
> meanrate=array(0,c(94,1))
> for (i in 1:94)
+ meanrate[i]=mean(rgamma(1000, y[i] + alpha, e[i] + alpha/mu))
> hospital[meanrate==min(meanrate)]
```

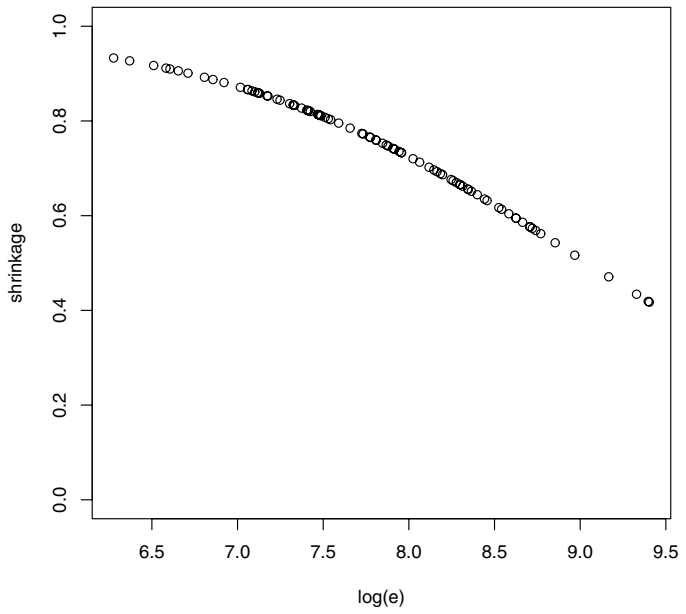

Fig. 7.10. Plot of the posterior shrinkages against the log exposures for the heart transplant example.

```
[1] 85
```

We identify hospital 85 as the one with the smallest true mortality rate.

Suppose we wish to compare hospital i with hospital j. One first obtains simulated draws from the marginal distribution of (λ_i, λ_j). Then the probability that hospital i has a smaller mortality rate, $\mathrm{P}(\lambda_i < \lambda_j)$, can be estimated by the proportion of simulated (λ_i, λ_j) pairs where λ_i is smaller than λ_j. In the following R code, we compute these probabilities for all pairs of hospitals and the results are stored in the matrix `better`. The probability that hospital i's rate is smaller than hospital j's rate is stored in the ith row and jth element of `better`.

```
> better=array(0,c(94,94))
> for (i in 1:94){
+   for (j in (i+1):94){
+   if (j <=94) {
+   lami=rgamma(1000,y[i]+alpha,e[i]+alpha/mu)
+   lamj=rgamma(1000,y[j]+alpha,e[j]+alpha/mu)
+   better[i,j]=mean(lami<lamj)
```

```
+   better[j,i]=1-better[i,j]
+ }}}
```

To compare the best hospital 85 with the remaining hospitals, we display the 85th column of the matrix `better`. These give the probabilities $P(\lambda_i < \lambda_{85})$ for all i. We display these probabilities for the first 24 hospitals. Note that hospital 85 is better than most of these hospitals since most of the posterior probabilities are close to zero.

```
> better[1:24,85]
```

```
 [1] 0.166 0.184 0.078 0.114 0.131 0.217 0.205 0.165 0.040 0.196
[11] 0.192 0.168 0.184 0.071 0.062 0.196 0.231 0.056 0.303 0.127
[21] 0.160 0.135 0.041 0.070
```

7.9 Posterior Predictive Model Checking

In Section 7.3, we used the posterior predictive distribution to examine the suitability of the "equal rates" model where $\lambda_1 = ... = \lambda_{94}$, and we saw that the model seemed inadequate in explaining the number of transplant deaths for individual hospitals. Here we use the same methodology to check the appropriateness of the exchangeable model.

Again we consider hospital 94, which experienced 17 deaths. Recall that simulated draws of the hyperparameters α and μ are contained in the vectors `alpha` and `mu`, respectively. To simulate from the predictive distribution of y_{94}^* we first simulate draws of the posterior density of λ_{94}

```
> lam94=rgamma(1000,y[94]+alpha,e[94]+alpha/mu)
```

and then simulate draws of y_{94}^* from a Poisson distribution with mean $e_{94}\lambda_{94}$.

```
> ys94=rpois(1000,e[94]*lam94)
```

Fig. 7.11 displays the histogram of y_{94}^* and places a vertical line on top corresponding to the value $y_{94} = 17$ using the commands

```
> hist(ys94,breaks=seq(1.5,39.5,by=1))
> lines(y[94]*c(1,1),c(0,100),lwd=3)
```

Note that in this case the observed number of deaths for this hospital is in the middle of the predictive distribution that indicates agreement of this observation with the fitted model.

Again this exchangeable model can check the consistency of the observed y_i with its posterior predictive distribution for all hospitals. In the following R code, we compute the probability that the future observation y_i^* is at least as extreme as y_i for all observations; the probabilities are placed in the vector `pout.exchange`.

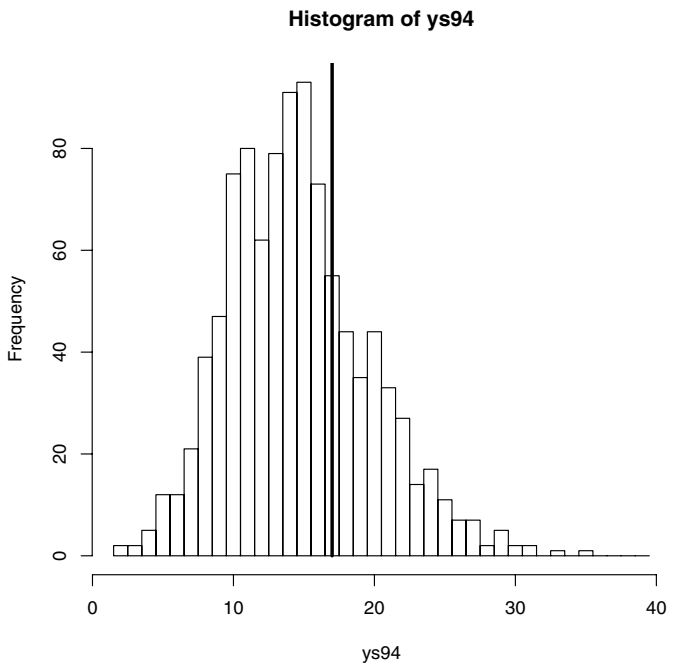

Fig. 7.11. Histogram of posterior predictive distribution of y^*_{94} for hospital 94 from the exchangeable model. The observed value of y_{94} is indicated by the vertical line.

```
> pout.exchange=0*y
> for (i in 1:94){
+   lami=rgamma(1000,y[i]+alpha,e[i]+alpha/mu)
+   ysi=rpois(1000,e[i]*lami)
+   pleft=sum(ysi<=y[i])/1000
+   pright=sum(ysi>=y[i])/1000
+   pout.exchange[i]=min(pleft,pright)
+ }
```

Recall that the probabilities of at least as extreme for the equal means model were contained in the vector `pout`. To compare the goodness of fits of the two models, Fig. 7.12 shows a scatterplot of the two sets of probabilities with a comparison line $y = x$ placed on top.

```
> plot(pout,pout.exchange,xlab="P(extreme), equal means",
+ ylab="P(extreme), exchangeable")
> abline(0,1)
```

Note that the probabilities of extreme for the exchangeable model are larger, indicating that the observations are more consistent with the exchangeable

fitted model. Note that only two of the observations have a probability smaller than 0.1 for the exchangeable model, indicating general agreement of the observed data with this model.

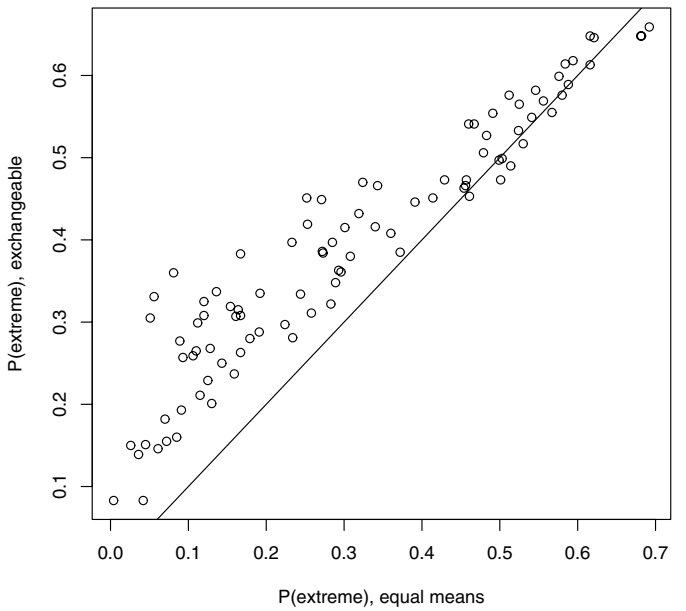

Fig. 7.12. Scatterplot of posterior predictive probabilities of "at least as extreme" for the equal means and exchangeable models.

7.10 Further Reading

Gelman et al (2003), chapter 5, provide a good introduction to hierarchical models. Carlin and Louis (2000), chapter 3, introduce hierarchical modeling from an empirical Bayes perspective. Posterior predictive model checking is described as a general method for model checking in chapter 6 of Gelman et al (2003). The use of hierarchical modeling to analyze the heart transplant data is described in Christiansen and Morris (1995).

7.11 Summary of R Functions

`poissgamexch` – computes the logarithm of the posterior for the parameters (log alpha, log mu) in a Poisson/gamma model
Usage: `poissgamexch(theta,datapar)`
Arguments: `theta`, matrix of parameter values where each row represents a value of (log alpha, log mu); `datapar`, list with components `data` (matrix with column of counts and column of exposures) and `z0`, the value of the second-stage hyperparameter
Value: vector of values of the log posterior where each value corresponds to each row of the parameters in theta

7.12 Exercises

1. **Normal/normal exchangeable model**
   Suppose we have J independent experiments, where in the jth experiment, we observe the single observation y_j that is normally distributed with mean θ_j and known variance σ_j^2. Suppose the parameters $\theta_1, ..., \theta_J$ are drawn from a normal population with mean μ and variance τ^2. The vector of hyperparameters (μ, τ) is assigned a uniform prior. Gelman et al (2003) describe the posterior calculations for this model. To summarize,
   - Conditional on the hyperparameters μ and τ, the θ_j have independent posterior distributions, where $\theta_j|\mu, \tau, y$ is normally distributed with mean $\hat{\theta}_j$ and variance V_j, where

$$\hat{\theta}_j = \frac{y_j/\sigma_j^2 + \mu/\tau^2}{1/\sigma_j^2 + 1/\tau^2},\ V_j = \frac{1}{1/\sigma_j^2 + 1/\tau^2}.$$

   - The marginal posterior density of the hyperparameters (μ, τ) is given by

$$g(\mu, \tau|y) \propto \prod_{j=1}^{J} \phi(y_j|\mu, \sqrt{\sigma_j^2 + \tau^2}),$$

     where $\phi(y|\mu, \sigma)$ denotes the normal density with mean μ and standard deviation σ.

   To illustrate this model, Gelman et al (2003) describe the results of independent experiments to determine the effects of special coaching programs on SAT scores. For the jth experiment, one observes an estimated coaching effect y_j with associated standard error σ_j; the values of the effects and standard errors are displayed in Table 7.1. The objective is to combine the coaching estimates in some way to obtain improved estimates at the true effects θ_j.

Table 7.1. Observed effects of special preparation on SAT scores in eight randomized experiments.

School	Treatment Effect y_j	Standard Error σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

a) Write an R function to compute the logarithm of the posterior density of the hyperparameters μ and $\log\tau$. (Don't forget to include the Jacobian term in the transformation to $(\mu, \log\tau)$.) Use a simulation algorithm such as Gibbs sampling (function `gibbs`), random walk Metropolis (function `rwmetrop`), or independence Metropolis (function `indepmetrop`) to obtain a sample of size 1000 from the posterior of $(\mu, \log\tau)$.

b) Using the simulated sample from the marginal posterior of $(\mu, \log\tau)$, simulate 1000 draws from the joint posterior density of the means $\theta_1, ..., \theta_J$. Summarize the posterior distribution of each θ_j by the computation of a posterior mean and posterior standard deviation.

2. **Normal/normal exchangeable model (continued)**
We assume that the sampling algorithm in Exercise 7.1 has been followed and one has simulated a sample of 1000 values from the marginal posterior of the hyperparameters μ and $\log\tau$, and also from the posterior densities of $\theta_1, ..., \theta_J$.

a) The posterior mean of θ_j, conditional on μ and τ, can be written as

$$E(\theta_j|y, \mu, \tau) = (1 - B_j)y_j + B_j\mu,$$

where $B_j = \tau^{-2}/(\tau^{-2} + \sigma_j^{-2})$ is the size of the shrinkage of y_j toward the mean μ. For all observations, compute the shrinkage size $E(B_j|y)$ from the simulated draws of the hyperparameters. Rank the schools from the largest shrinkage to the smallest shrinkage and explain why there are differences.

b) School A had the largest observed coaching effect of 28. From the simulated draws from the joint distribution of $\theta_1, ..., \theta_J$, compute the posterior probability $\mathrm{P}(\theta_1 > \theta_j)$ for $j = 2, ..., J$.

3. **Binomial/beta exchangeable model**
In Chapter 5, we described the problem of simultaneously estimating the rates of death from stomach cancer for males at risk in the age

bracket 45–64 for the largest cities in Missouri. The dataset is available as `cancermortality` in the LearnBayes package. Assume that the numbers of cancer deaths $\{y_j\}$ are independent, where y_j is binomial with sample size n_j and probability of death p_j. To model a prior belief of exchangeability, it is assumed that $p_1, ..., p_{20}$ are a random sample from a beta distribution with parameters a and b. We reparameterize the beta parameters a and b to new values

$$\eta = \frac{a}{a+b},\ K = a + b.$$

The hyperparameter η is the prior mean of each p_j and K is a precision parameter. At the last stage of this model, we assign (η, K) the noninformative proper prior

$$g(\eta, K) = \frac{1}{(1+K)^2},\ 0 < \eta < 1, K > 0.$$

Due to the conjugate form of the prior, one can derive the following posterior distributions.

- Conditional on the values of the hyperparameters η and K, the probabilities $p_1, ..., p_{20}$ are independent, with p_j distributed beta with parameters $a_j = K\eta + y_j$ and $b_j = K(1-\eta) + n_j - y_j$.
- The marginal posterior density of (η, K) has the form

$$g(\eta, K|y) \propto \frac{1}{(1+K)^2} \prod_{j=1}^{20} \frac{B(K\eta + y_j, K(1-\eta) + n_j - y_j)}{B(K\eta, K(1-\eta)},$$

where $K > 0$ and $0 < \eta < 1$.

a) To summarize the posterior distribution of the hyperparameters η and K, first transform the parameters to the real line by the reexpressions $\theta_1 = \log K$ and $\theta_2 = \log(\eta/(1-\eta))$. Write an R function to compute values of the log posterior of θ_1 and θ_2.

b) Use a simulation algorithm such as Gibbs sampling (function `gibbs`), random walk Metropolis (function `rwmetrop`), or independence Metropolis (function `indepmetrop`) to obtain a sample of size 1000 from the posterior of (θ_1, θ_2). Summarize the posterior distributions of K and η by 90% interval estimates.

c) Using the simulated sample from the marginal posterior of (θ_1, θ_2), simulate 1000 draws from the joint posterior density of the probabilities $p_1, ..., p_{20}$. Summarize the posterior distribution of each p_j by a 90% interval estimate.

4. **Binomial/beta exchangeable model (continued)**
We assume that the sampling algorithm in Exercise 7.3 has been followed and one has simulated a sample of 1000 values from the marginal posterior of the hyperparameters K and m, and also from the posterior densities of $p_1, ..., p_{20}$.

a) Let y_j^* denote the number of cancer deaths of a future sample of size n_j from the jth city in Missouri. Conditional on the probability p_j distribution of y_j^* is binomial(n_j, p_j). For city 1 (with $n_j = 1083$ patients) and city 15 (with $n_j = 53637$ patients), simulate a sample of 1000 values from the posterior predictive distribution of y_j^*.

b) For cities 1 and 15, the observed numbers of cancer deaths were 0 and 54, respectively. By comparing the observed values of y_j against the respective predictive distributions, decide if these values are consistent with the binomial/beta exchangeable model.