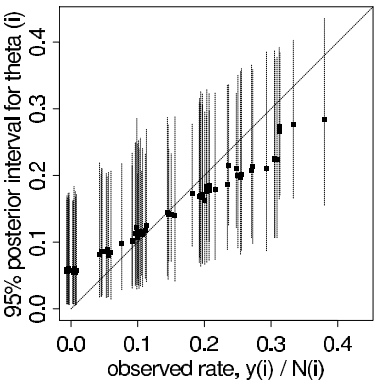

Figure 5.4 *Posterior medians and 95% intervals of rat tumor rates, θ_j (plotted vs. observed tumor rates y_j/n_j), based on simulations from the joint posterior distribution. The 45° line corresponds to the unpooled estimates, $\hat{\theta}_i = y_i/n_i$. The horizontal positions of the line have been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

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,

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

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

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

with sampling variance

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

We can then write the likelihood for each θ_j using the sufficient statistics, $\overline{y}_{.j}$:

$$\overline{y}_{.j}|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2), \tag{5.12}$$

a notation that will prove useful later because of the flexibility in allowing a separate variance σ_j^2 for the mean of each group j. For the rest of this chapter, all expressions will be implicitly conditional on the known values σ_j^2. The problem of estimating a set of means with unknown variances will require some additional computational methods, presented in Sections 11.6 and 13.6. Although rarely strictly true, the assumption of known variances at the sampling level of the model is often an adequate approximation.

The treatment of the model provided in this section is also appropriate for situations in which the variances differ for reasons other than the number of data points in the experiment. In fact, the likelihood (5.12) can appear in much more general contexts than that stated here. For example, if the group sizes n_j are large enough, then the means $\overline{y}_{.j}$ are approximately normally distributed, given θ_j, even when the data y_{ij} are not. Other applications where the actual likelihood is well approximated by (5.12) appear in the next two sections.

Constructing a prior distribution from pragmatic considerations

Rather than considering immediately the problem of specifying a prior distribution for the parameter vector $\theta = (\theta_1, \ldots, \theta_J)$, let us consider what sorts of posterior estimates might be reasonable for θ, given data (y_{ij}). A simple natural approach is to estimate θ_j by $\overline{y}_{.j}$, the average outcome in experiment j. But what if, for example, there are $J = 20$ experiments with only $n_j = 2$ observations per experimental group, and the groups are 20 pairs of assays taken from the same strain of rat, under essentially identical conditions? The two observations per group do not permit accurate estimates. Since the 20 groups are from the same strain of rat, we might now prefer to estimate each θ_j by the pooled estimate,

$$\overline{y}_{..} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \overline{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}. \tag{5.13}$$

To decide which estimate to use, a traditional approach from classical statistics is to perform an analysis of variance F test for differences among means: if the J group means appear significantly variable, choose separate sample means, and if the variance between the group means is not significantly greater than what could be explained by individual variability within groups, use $\overline{y}_{..}$. The theoretical analysis of variance table is as follows, where τ^2 is the variance of $\theta_1, \ldots, \theta_J$. For simplicity, we present the analysis of variance for a balanced design in which $n_j = n$ and $\sigma_j^2 = \sigma^2/n$ for all j.

	df	SS	MS	$\mathrm{E}(\mathrm{MS}\|\sigma^2, \tau)$
Between groups	$J-1$	$\sum_i \sum_j (\overline{y}_{.j} - \overline{y}_{..})^2$	$\mathrm{SS}/(J-1)$	$n\tau^2 + \sigma^2$
Within groups	$J(n-1)$	$\sum_i \sum_j (y_{ij} - \overline{y}_{.j})^2$	$\mathrm{SS}/(J(n-1))$	σ^2
Total	$Jn-1$	$\sum_i \sum_j (y_{ij} - \overline{y}_{..})^2$	$\mathrm{SS}/(Jn-1)$	

In the classical random-effects analysis of variance, one computes the sum of squares (SS) and the mean square (MS) columns of the table and uses the 'between' and 'within' mean squares to estimate τ. If the ratio of between to within mean squares is significantly greater than 1, then the analysis of variance suggests separate estimates, $\hat{\theta}_j = \overline{y}_{.j}$ for each j. If the ratio of mean squares is not 'statistically significant,' then the F test cannot 'reject the hypothesis' that $\tau = 0$, and pooling is reasonable: $\hat{\theta}_j = \overline{y}_{..}$, for all j. We discuss Bayesian analysis of variance in Section 15.6 in the context of hierarchical regression models.

But we are not forced to choose between complete pooling and none at all. An alternative is to use a weighted combination:

$$\hat{\theta}_j = \lambda_j \overline{y}_{.j} + (1 - \lambda_j)\overline{y}_{..},$$

where λ_j is between 0 and 1.

What kind of prior models produce these various posterior estimates?

1. The unpooled estimate $\hat{\theta}_j = \overline{y}_{.j}$ is the posterior mean if the J values θ_j have independent uniform prior densities on $(-\infty, \infty)$.
2. The pooled estimate $\hat{\theta} = \overline{y}_{..}$ is the posterior mean if the J values θ_j are restricted to be equal, with a uniform prior density on the common θ.
3. The weighted combination is the posterior mean if the J values θ_j have independent and identically distributed normal prior densities.

All three of these options are exchangeable in the θ_j's, and options 1 and 2 are special cases of option 3. No pooling corresponds to $\lambda_j \equiv 1$ for all j and an infinite prior variance for the θ_j's, and complete pooling corresponds to $\lambda_j \equiv 0$ for all j and a zero prior variance for the θ_j's.

The hierarchical model

For the convenience of conjugacy (more accurately, partial conjugacy), we assume that the parameters θ_j are drawn from a normal distribution with hyperparameters (μ, τ):

$$\begin{aligned} p(\theta_1, \ldots, \theta_J | \mu, \tau) &= \prod_{j=1}^{J} \mathrm{N}(\theta_j | \mu, \tau^2) \qquad (5.14) \\ p(\theta_1, \ldots, \theta_J) &= \int \prod_{j=1}^{J} \left[\mathrm{N}(\theta_j | \mu, \tau^2)\right] p(\mu, \tau) d(\mu, \tau). \end{aligned}$$

That is, the θ_j's are conditionally independent given (μ, τ). The hierarchical model also permits the interpretation of the θ_j's as a random sample from a shared population distribution, as illustrated in Figure 5.1 for the rat tumors.

We assign a noninformative uniform hyperprior distribution to μ, given τ:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau). \qquad (5.15)$$

The uniform prior density for μ is generally reasonable for this problem; because the combined data from all J experiments are generally highly informative about μ, we can afford to be vague about its prior distribution. We defer discussion of the prior distribution of τ to later in the analysis, although relevant principles have already been discussed in the context of the rat tumor example. As usual, we first work out the answer conditional on the hyperparameters and then consider their prior and posterior distributions.

The joint posterior distribution

Combining the sampling model for the observable y_{ij}'s and the prior distribution yields the joint posterior distribution of all the parameters and hyperparameters, which we can express in terms of the sufficient statistics, $\overline{y}_{.j}$:

$$\begin{aligned} p(\theta, \mu, \tau | y) &\propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta) \\ &\propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j} | \theta_j, \sigma_j^2), \qquad (5.16) \end{aligned}$$

where we can ignore factors that depend only on y and the parameters σ_j, which are assumed known for this analysis.

The conditional posterior distribution of the normal means, given the hyperparameters

As in the general hierarchical structure, the parameters θ_j are independent in the prior distribution (given μ and τ) and appear in different factors in the likelihood (5.11); thus, the conditional posterior distribution $p(\theta|\mu,\tau,y)$ factors into J components.

Conditional on the hyperparameters, we simply have J independent unknown normal means, given normal prior distributions, so we can use the methods of Section 2.5 independently on each θ_j. The conditional posterior distributions for the θ_j's are independent, and

$$\theta_j|\mu,\tau,y \sim \mathrm{N}(\hat{\theta}_j, V_j),$$

where

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\overline{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}. \tag{5.17}$$

The posterior mean is a precision-weighted average of the prior population mean and the sample mean of the jth group; these expressions for $\hat{\theta}_j$ and V_j are functions of μ and τ as well as the data. The conditional posterior density for each θ_j given μ,τ is proper.

The marginal posterior distribution of the hyperparameters

The solution so far is only partial because it depends on the unknown μ and τ. The next step in our approach is a full Bayesian treatment for the hyperparameters. Section 5.3 mentions integration or analytic computation as two approaches for obtaining $p(\mu,\tau|y)$ from the joint posterior density $p(\theta,\mu,\tau|y)$. For the hierarchical normal model, we can simply consider the information supplied by the data about the hyperparameters directly:

$$p(\mu,\tau|y) \propto p(\mu,\tau)p(y|\mu,\tau).$$

For many problems, this decomposition is no help, because the 'marginal likelihood' factor, $p(y|\mu,\tau)$, cannot generally be written in closed form. For the normal distribution, however, the marginal likelihood has a particularly simple form. The marginal distributions of the group means $\overline{y}_{.j}$, averaging over θ, are independent (but not identically distributed) normal:

$$\overline{y}_{.j}|\mu,\tau \sim \mathrm{N}(\mu, \sigma_j^2 + \tau^2).$$

Thus we can write the marginal posterior density as

$$p(\mu,\tau|y) \propto p(\mu,\tau)\prod_{j=1}^{J}\mathrm{N}(\overline{y}_{.j}|\mu,\sigma_j^2+\tau^2). \tag{5.18}$$

Posterior distribution of μ given τ. We could use (5.18) to compute directly the posterior distribution $p(\mu,\tau|y)$ as a function of two variables and proceed as in the rat tumor example. For the normal model, however, we can further simplify by integrating over μ, leaving a simple univariate numerical computation of $p(\tau|y)$. We factor the marginal posterior density of the hyperparameters as we did the prior density (5.15):

$$p(\mu,\tau|y) = p(\mu|\tau,y)p(\tau|y). \tag{5.19}$$

The first factor on the right side of (5.19) is just the posterior distribution of μ if τ were known. From inspection of (5.18) with τ assumed known, and with a uniform conditional

prior density $p(\mu|\tau)$, the log posterior distribution is found to be quadratic in μ; thus, $p(\mu|\tau, y)$ must be normal. The mean and variance of this distribution can be obtained immediately by considering the group means $\overline{y}_{.j}$ as J independent estimates of μ with variances $(\sigma_j^2 + \tau^2)$. Combining the data with the uniform prior density $p(\mu|\tau)$ yields

$$\mu|\tau, y \sim \mathrm{N}(\hat{\mu}, V_\mu),$$

where $\hat{\mu}$ is the precision-weighted average of the $\overline{y}_{.j}$-values, and V_μ^{-1} is the total precision:

$$\hat{\mu} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}\overline{y}_{.j}}{\sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}. \tag{5.20}$$

The result is a proper posterior density for μ, given τ.

Posterior distribution of τ. We can now obtain the posterior distribution of τ analytically from (5.19) and substitution of (5.18) and (5.20) for the numerator and denominator, respectively:

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)} \\ &\propto \frac{p(\tau)\prod_{j=1}^J \mathrm{N}(\overline{y}_{.j}|\mu, \sigma_j^2 + \tau^2)}{\mathrm{N}(\mu|\hat{\mu}, V_\mu)}. \end{aligned}$$

This identity must hold for any value of μ (in other words, all the factors of μ must cancel when the expression is simplified); in particular, it holds if we set μ to $\hat{\mu}$, which makes evaluation of the expression simple:

$$\begin{aligned} p(\tau|y) &\propto \frac{p(\tau)\prod_{j=1}^J \mathrm{N}(\overline{y}_{.j}|\hat{\mu}, \sigma_j^2 + \tau^2)}{\mathrm{N}(\hat{\mu}|\hat{\mu}, V_\mu)} \\ &\propto p(\tau)V_\mu^{1/2}\prod_{j=1}^J (\sigma_j^2 + \tau^2)^{-1/2}\exp\left(-\frac{(\overline{y}_{.j} - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right), \end{aligned} \tag{5.21}$$

with $\hat{\mu}$ and V_μ defined in (5.20). Both expressions are functions of τ, which means that $p(\tau|y)$ is a complicated function of τ.

Prior distribution for τ. To complete our analysis, we must assign a prior distribution to τ. For convenience, we use a diffuse noninformative prior density for τ and hence must examine the resulting posterior density to ensure it has a finite integral. For our illustrative analysis, we use the uniform prior distribution, $p(\tau) \propto 1$. We leave it as an exercise to show mathematically that the uniform prior density for τ yields a proper posterior density and that, in contrast, the seemingly reasonable 'noninformative' prior distribution for a variance component, $p(\log\tau) \propto 1$, yields an improper posterior distribution for τ. Alternatively, in applications it involves little extra effort to determine a 'best guess' and an upper bound for the population variance τ, and a reasonable prior distribution can then be constructed from the scaled inverse-χ^2 family (the natural choice for variance parameters), matching the 'best guess' to the mean of the scaled inverse-χ^2 density and the upper bound to an upper percentile such as the 99th. Once an initial analysis is performed using the noninformative 'uniform' prior density, a sensitivity analysis with a more realistic prior distribution is often desirable.

Computation

For this model, computation of the posterior distribution of θ is most conveniently performed via simulation, following the factorization used above:

$$p(\theta, \mu, \tau|y) = p(\tau|y)p(\mu|\tau, y)p(\theta|\mu, \tau, y).$$

The first step, simulating τ, is easily performed numerically using the inverse cdf method (see Section 1.9) on a grid of uniformly spaced values of τ, with $p(\tau|y)$ computed from (5.21). The second and third steps, simulating μ and then θ, can both be done easily by sampling from normal distributions, first (5.20) to obtain μ and then (5.17) to obtain the θ_j's independently.

Posterior predictive distributions

Sampling from the posterior predictive distribution of new data, either from a current or new batch, is straightforward given draws from the posterior distribution of the parameters. We consider two scenarios: (1) future data $\tilde{y}$ from the current set of batches, with means $\theta = (\theta_1, \ldots, \theta_J)$, and (2) future data $\tilde{y}$ from $\tilde{J}$ future batches, with means $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{\tilde{J}})$. In the latter case, we must also specify the $\tilde{J}$ individual sample sizes $\tilde{n}_j$ for the future batches.

To obtain a draw from the posterior predictive distribution of new data $\tilde{y}$ from the current batch of parameters, θ, first obtain a draw from $p(\theta, \mu, \tau|y)$ and then draw the predictive data $\tilde{y}$ from (5.11).

To obtain posterior predictive simulations of new data $\tilde{y}$ for $\tilde{J}$ new groups, perform the following three steps: first, draw (μ, τ) from their posterior distribution; second, draw $\tilde{J}$ new parameters $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{\tilde{J}})$ from the population distribution $p(\tilde{\theta}_j|\mu, \tau)$, which is the population, or prior, distribution for θ given the hyperparameters (equation (5.14)); and third, draw $\tilde{y}$ given $\tilde{\theta}$ from the data distribution (5.11).

Difficulty with a natural non-Bayesian estimate of the hyperparameters

To see some advantages of our fully Bayesian approach, we compare it to an approximate method that is sometimes used based on a *point estimate* of μ and τ from the data. Unbiased point estimates, derived from the analysis of variance presented earlier, are

$$\begin{aligned} \hat{\mu} &= \overline{y}_{\cdot\cdot} \\ \hat{\tau}^2 &= (\mathrm{MS}_B - \mathrm{MS}_W)/n. \end{aligned} \tag{5.22}$$

The terms MS_B and MS_W are the 'between' and 'within' mean squares, respectively, from the analysis of variance. In this alternative approach, inference for $\theta_1, \ldots, \theta_J$ is based on the conditional posterior distribution, $p(\theta|\hat{\mu}, \hat{\tau})$, given the point estimates.

As we saw in the rat tumor example of the previous section, the main problem with substituting point estimates for the hyperparameters is that it ignores our real uncertainty about them. The resulting inference for θ cannot be interpreted as a Bayesian posterior summary. In addition, the estimate $\hat{\tau}^2$ in (5.22) has the flaw that it can be negative! The problem of a negative estimate for a variance component can be avoided by setting $\hat{\tau}^2$ to zero in the case that MS_W exceeds MS_B, but this creates new issues. Estimating $\tau^2 = 0$ whenever $\mathrm{MS}_W > \mathrm{MS}_B$ seems too strong a claim: if $\mathrm{MS}_W > \mathrm{MS}_B$, then the sample size is too small for τ^2 to be distinguished from zero, but this is not the same as saying we know that $\tau^2 = 0$. The latter claim, made implicitly by the point estimate, implies that all the group means θ_j are absolutely identical, which leads to scientifically indefensible claims, as we shall see in the example in the next section. It is possible to construct a point estimate

of (μ, τ) to avoid this particular difficulty, but it would still have the problem, common to all point estimates, of ignoring uncertainty.

5.5 Example: parallel experiments in eight schools

We illustrate the hierarchical normal model with a problem in which the Bayesian analysis gives conclusions that differ in important respects from other methods.

A study was performed for the Educational Testing Service to analyze the effects of special coaching programs on test scores. Separate randomized experiments were performed to estimate the effects of coaching programs for the SAT-V (Scholastic Aptitude Test-Verbal) in each of eight high schools. The outcome variable in each study was the score on a special administration of the SAT-V, a standardized multiple choice test administered by the Educational Testing Service and used to help colleges make admissions decisions; the scores can vary between 200 and 800, with mean about 500 and standard deviation about 100. The SAT examinations are designed to be resistant to short-term efforts directed specifically toward improving performance on the test; instead they are designed to reflect knowledge acquired and abilities developed over many years of education. Nevertheless, each of the eight schools in this study considered its short-term coaching program to be successful at increasing SAT scores. Also, there was no prior reason to believe that any of the eight programs was more effective than any other or that some were more similar in effect to each other than to any other.

The results of the experiments are summarized in Table 5.2. All students in the experiments had already taken the PSAT (Preliminary SAT), and allowance was made for differences in the PSAT-M (Mathematics) and PSAT-V test scores between coached and uncoached students. In particular, in each school the estimated coaching effect and its standard error were obtained by an analysis of covariance adjustment (that is, a linear regression was performed of SAT-V on treatment group, using PSAT-M and PSAT-V as control variables) appropriate for a completely randomized experiment. A separate regression was estimated for each school. Although not simple sample means (because of the covariance adjustments), the estimated coaching effects, which we label y_j, and their sampling variances, σ_j^2, play the same role in our model as $\overline{y}_{.j}$ and σ_j^2 in the previous section. The estimates y_j are obtained by independent experiments and have approximately normal sampling distributions with sampling variances that are known, for all practical purposes, because the sample sizes in all of the eight experiments were relatively large, over thirty students in each school (recall the discussion of data reduction in Section 4.1). Incidentally, an increase of eight points on the SAT-V corresponds to about one more test item correct.

Inferences based on nonhierarchical models and their problems

Before fitting the hierarchical Bayesian model, we first consider two simpler nonhierarchical methods—estimating the effects from the eight experiments independently, and complete pooling—and discuss why neither of these approaches is adequate for this example.

Separate estimates. A cursory examination of Table 5.2 may at first suggest that some coaching programs have moderate effects (in the range 18–28 points), most have small effects (0–12 points), and two have small negative effects; however, when we take note of the standard errors of these estimated effects, we see that it is difficult statistically to distinguish between any of the experiments. For example, treating each experiment separately and applying the simple normal analysis in each yields 95% posterior intervals that all overlap substantially.

A pooled estimate. The general overlap in the posterior intervals based on independent analyses suggests that all experiments might be estimating the same quantity. Under the

School	Estimated treatment effect, y_j	Standard error of effect estimate, 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

Table 5.2 *Observed effects of special preparation on SAT-V scores in eight randomized experiments. Estimates are based on separate analyses for the eight experiments.*

hypothesis that all experiments have the same effect and produce independent estimates of this common effect, we could treat the data in Table 5.2 as eight normally distributed observations with known variances. With a noninformative prior distribution, the posterior mean for the common coaching effect in the schools is $\overline{y}_{\cdot\cdot}$, as defined in equation (5.13) with y_j in place of $\overline{y}_{\cdot j}$. This pooled estimate is 7.7, and the posterior variance is $(\sum_{j=1}^{8} \frac{1}{\sigma_j^2})^{-1} = 16.6$ because the eight experiments are independent. Thus, we would estimate the common effect to be 7.7 points with standard error equal to $\sqrt{16.6} = 4.1$, which would lead to the 95% posterior interval $[-0.5, 15.9]$, or approximately $[8 \pm 8]$. Supporting this analysis, the classical test of the hypothesis that all θ_j's are estimating the same quantity yields a χ^2 statistic less than its degrees of freedom (seven, in this case): $\sum_{j=1}^{8}(y_j - \overline{y}_{\cdot\cdot})^2/\sigma_i^2 = 4.6$. To put it another way, the estimate $\hat{\tau}^2$ from (5.22) is negative.

Would it be possible to have one school's observed effect be 28 just by chance, if the coaching effects in all eight schools were really the same? To get a feeling for the natural variation that we would expect across eight studies if this assumption were true, suppose the estimated treatment effects are eight independent draws from a normal distribution with mean 8 points and standard deviation 13 points (the square root of the mean of the eight variances σ_j^2). Then, based on the expected values of normal order statistics, we would expect the largest observed value of y_j to be about 26 points and the others, in diminishing order, to be about 19, 14, 10, 6, 2, −3, and −9 points. These expected effect sizes are consistent with the set of observed effect sizes in Table 5.2. Thus, it would appear imprudent to believe that school A really has an effect as large as 28 points.

Difficulties with the separate and pooled estimates. To see the problems with the two extreme attitudes—the separate analyses that consider each θ_j separately, and the alternative view (a single common effect) that leads to the pooled estimate—consider θ_1, the effect in school A. The effect in school A is estimated as 28.4 with a standard error of 14.9 under the separate analysis, versus a pooled estimate of 7.7 with a standard error of 4.1 under the common-effect model. The separate analyses of the eight schools imply the following posterior statement: 'the probability is $\frac{1}{2}$ that the true effect in A is more than 28.4,' a doubtful statement, considering the results for the other seven schools. On the other hand, the pooled model implies the following statement: 'the probability is $\frac{1}{2}$ that the true effect in A is less than 7.7,' which, despite the non-significant χ^2 test, seems an inaccurate summary of our knowledge. The pooled model also implies the statement: 'the probability is $\frac{1}{2}$ that the true effect in A is less than the true effect in C,' which also is difficult to justify given the data in Table 5.2. As in the theoretical discussion of the previous section, neither estimate is fully satisfactory, and we would like a compromise that combines information

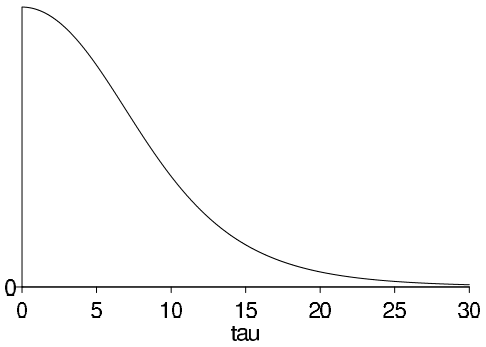

Figure 5.5 *Marginal posterior density, $p(\tau|y)$, for standard deviation of the population of school effects θ_j in the educational testing example.*

from all eight experiments without assuming all the θ_j's to be equal. The Bayesian analysis under the hierarchical model provides exactly that.

Posterior simulation under the hierarchical model

Consequently, we compute the posterior distribution of $\theta_1, \ldots, \theta_8$, based on the normal model presented in Section 5.4. (More discussion of the reasonableness of applying this model in this problem appears in Sections 6.5 and 17.4.) We draw from the posterior distribution for the Bayesian model by simulating the random variables τ, μ, and θ, in that order, from their posterior distribution, as discussed at the end of the previous section. The sampling standard deviations, σ_j, are assumed known and equal to the values in Table 5.2, and we assume independent uniform prior densities on μ and τ.

Results

The marginal posterior density function, $p(\tau|y)$ from (5.21), is plotted in Figure 5.5. Values of τ near zero are most plausible; zero is the most likely value, values of τ larger than 10 are less than half as likely as $\tau = 0$, and $\Pr(\tau > 25) \approx 0$. Inference regarding the marginal distributions of the other model parameters and the joint distribution are obtained from the simulated values. Illustrations are provided in the discussion that follows this section. In the normal hierarchical model, however, we learn a great deal by considering the conditional posterior distributions given τ (and averaged over μ).

The conditional posterior means $\mathrm{E}(\theta_j|\tau, y)$ (averaging over μ) are displayed as functions of τ in Figure 5.6; the vertical axis displays the scale for the θ_j's. Comparing Figure 5.6 to Figure 5.5, which has the same scale on the horizontal axis, we see that for most of the likely values of τ, the estimated effects are relatively close together; as τ becomes larger, corresponding to more variability among schools, the estimates become more like the raw values in Table 5.2.

The lines in Figure 5.7 show the conditional standard deviations, $\mathrm{sd}(\theta_j|\tau, y)$, as a function of τ. As τ increases, the population distribution allows the eight effects to be more different from each other, and hence the posterior uncertainty in each individual θ_j increases, approaching the standard deviations in Table 5.2 in the limit of $\tau \to \infty$. (The posterior means and standard deviations for the components θ_j, given τ, are computed using the mean and variance formulas (2.7) and (2.8), averaging over μ; see Exercise 5.12.)

The general conclusion from an examination of Figures 5.5–5.7 is that an effect as large as 28.4 points in any school is unlikely. For the likely values of τ, the estimates in all schools are substantially less than 28 points. For example, even at $\tau = 10$, the probability

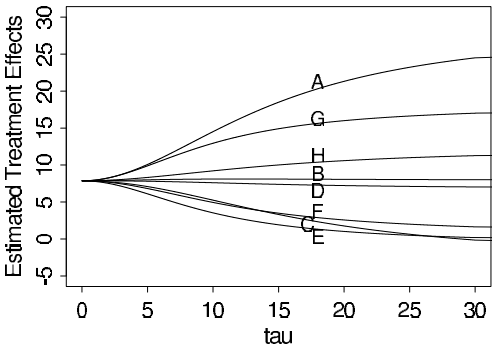

Figure 5.6 *Conditional posterior means of treatment effects, $E(\theta_j|\tau, y)$, as functions of the between-school standard deviation τ, for the educational testing example. The line for school* C *crosses the lines for* E *and* F *because* C *has a higher measurement error (see Table 5.2) and its estimate is therefore shrunk more strongly toward the overall mean in the Bayesian analysis.*

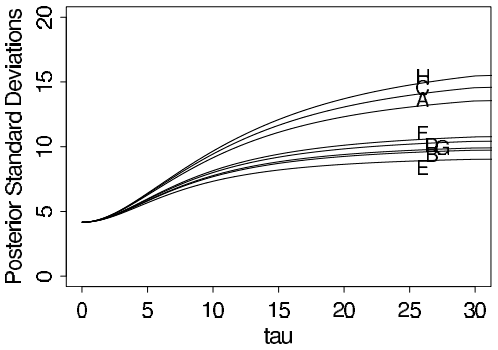

Figure 5.7 *Conditional posterior standard deviations of treatment effects,* $\mathrm{sd}(\theta_j|\tau, y)$*, as functions of the between-school standard deviation τ, for the educational testing example.*

that the effect in school A is less than 28 points is $\Phi[(28 - 14.5)/9.1] = 93\%$, where Φ is the standard normal cumulative distribution function; the corresponding probabilities for the effects being less than 28 points in the other schools are 99.5%, 99.2%, 98.5%, 99.96%, 99.8%, 97%, and 98%.

Of substantial importance, we do not obtain an accurate summary of the data if we condition on the posterior mode of τ. The technique of conditioning on a modal value (for example, the maximum likelihood estimate) of a hyperparameter such as τ is often used in practice (at least as an approximation), but it ignores the uncertainty conveyed by the posterior distribution of the hyperparameter. At $\tau = 0$, the inference is that all experiments have the same size effect, 7.7 points, and the same standard error, 4.1 points. Figures 5.5–5.7 certainly suggest that this answer represents too much pulling together of the estimates in the eight schools. The problem is especially acute in this example because the posterior mode of τ is on the boundary of its parameter space. A joint posterior modal estimate of $(\theta_1, \ldots, \theta_J, \mu, \tau)$ suffers from even worse problems in general.

Discussion

Table 5.3 summarizes the 200 simulated effect estimates for all eight schools. In one sense, these results are similar to the pooled 95% interval $[8 \pm 8]$, in that the eight Bayesian 95% intervals largely overlap and are median-centered between 5 and 10. In a second sense,

School	Posterior quantiles				
	2.5%	25%	median	75%	97.5%
A	−2	7	10	16	31
B	−5	3	8	12	23
C	−11	2	7	11	19
D	−7	4	8	11	21
E	−9	1	5	10	18
F	−7	2	6	10	28
G	−1	7	10	15	26
H	−6	3	8	13	33

Table 5.3: *Summary of 200 simulations of the treatment effects in the eight schools.*

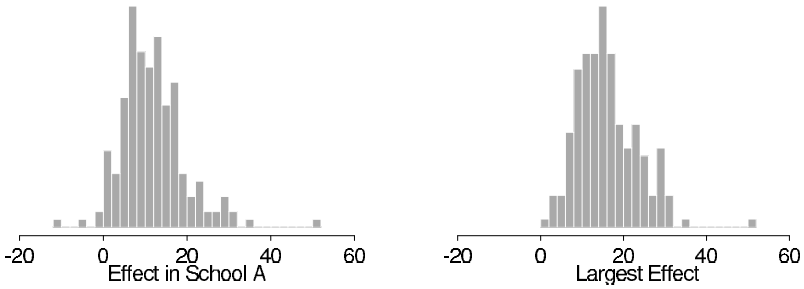

Figure 5.8 *Histograms of two quantities of interest computed from the 200 simulation draws: (a) the effect in school A, θ_1; (b) the largest effect,* $\max\{\theta_j\}$. *The jaggedness of the histograms is just an artifact caused by sampling variability from using only 200 random draws.*

the results in the table differ from the pooled estimate in a direction toward the eight independent answers: the 95% Bayesian intervals are each almost twice as wide as the one common interval and suggest substantially greater probabilities of effects larger than 16 points, especially in school A, and greater probabilities of negative effects, especially in school C. If greater precision were required in the posterior intervals, one could simulate more simulation draws; we use only 200 draws here to illustrate that a small simulation gives adequate inference for many practical purposes.

The ordering of the effects in the eight schools as suggested by Table 5.3 is essentially the same as would be obtained by the eight separate estimates. However, there are differences in the details; for example, the Bayesian probability that the effect in school A is as large as 28 points is less than 10%, which is substantially less than the 50% probability based on the separate estimate for school A.

As an illustration of the simulation-based posterior results, 200 simulations of school A's effect are shown in Figure 5.8a. Having simulated the parameter θ, it is easy to ask more complicated questions of this model. For example, what is the posterior distribution of $\max\{\theta_j\}$, the effect of the most successful of the eight coaching programs? Figure 5.8b displays a histogram of 200 values from this posterior distribution and shows that only 22 draws are larger than 28.4; thus, $\Pr(\max\{\theta_j\} > 28.4) \approx \frac{22}{200}$. Since Figure 5.8a gives the marginal posterior distribution of the effect in school A, and Figure 5.8b gives the marginal posterior distribution of the largest effect no matter which school it is in, the latter figure has larger values. For another example, we can estimate $\Pr(\theta_1 > \theta_3|y)$, the posterior probability that the coaching program is more effective in school A than in school C, by the proportion of simulated draws of θ for which $\theta_1 > \theta_3$; the result is $\frac{141}{200} = 0.705$.

To sum up, the Bayesian analysis of this example not only allows straightforward inferences about many parameters that may be of interest, but the hierarchical model is flexible

Study,	Raw data (deaths/total)		Log-odds,	sd,	Posterior quantiles of effect θ_j normal approx. (on log-odds scale)				
j	Control	Treated	y_j	σ_j	2.5%	25%	median	75%	97.5%
1	3/39	3/38	0.028	0.850	−0.57	−0.33	−0.24	−0.16	0.12
2	14/116	7/114	−0.741	0.483	−0.64	−0.37	−0.28	−0.20	−0.00
3	11/93	5/69	−0.541	0.565	−0.60	−0.35	−0.26	−0.18	0.05
4	127/1520	102/1533	−0.246	0.138	−0.45	−0.31	−0.25	−0.19	−0.05
5	27/365	28/355	0.069	0.281	−0.43	−0.28	−0.21	−0.11	0.15
6	6/52	4/59	−0.584	0.676	−0.62	−0.35	−0.26	−0.18	0.05
7	152/939	98/945	−0.512	0.139	−0.61	−0.43	−0.36	−0.28	−0.17
8	48/471	60/632	−0.079	0.204	−0.43	−0.28	−0.21	−0.13	0.08
9	37/282	25/278	−0.424	0.274	−0.58	−0.36	−0.28	−0.20	−0.02
10	188/1921	138/1916	−0.335	0.117	−0.48	−0.35	−0.29	−0.23	−0.13
11	52/583	64/873	−0.213	0.195	−0.48	−0.31	−0.24	−0.17	0.01
12	47/266	45/263	−0.039	0.229	−0.43	−0.28	−0.21	−0.12	0.11
13	16/293	9/291	−0.593	0.425	−0.63	−0.36	−0.28	−0.20	0.01
14	45/883	57/858	0.282	0.205	−0.34	−0.22	−0.12	0.00	0.27
15	31/147	25/154	−0.321	0.298	−0.56	−0.34	−0.26	−0.19	0.01
16	38/213	33/207	−0.135	0.261	−0.48	−0.30	−0.23	−0.15	0.08
17	12/122	28/251	0.141	0.364	−0.47	−0.29	−0.21	−0.12	0.17
18	6/154	8/151	0.322	0.553	−0.51	−0.30	−0.23	−0.13	0.15
19	3/134	6/174	0.444	0.717	−0.53	−0.31	−0.23	−0.14	0.15
20	40/218	32/209	−0.218	0.260	−0.50	−0.32	−0.25	−0.17	0.04
21	43/364	27/391	−0.591	0.257	−0.64	−0.40	−0.31	−0.23	−0.09
22	39/674	22/680	−0.608	0.272	−0.65	−0.40	−0.31	−0.23	−0.07

Table 5.4 *Results of 22 clinical trials of beta-blockers for reducing mortality after myocardial infarction, with empirical log-odds and approximate sampling variances. Data from Yusuf et al. (1985). Posterior quantiles of treatment effects are based on 5000 draws from a Bayesian hierarchical model described here. Negative effects correspond to reduced probability of death under the treatment.*

enough to adapt to the data, thereby providing posterior inferences that account for the partial pooling as well as the uncertainty in the hyperparameters.

5.6 Hierarchical modeling applied to a meta-analysis

Meta-analysis is an increasingly popular and important process of summarizing and integrating the findings of research studies in a particular area. As a method for combining information from several parallel data sources, meta-analysis is closely connected to hierarchical modeling. In this section we consider a relatively simple application of hierarchical modeling to a meta-analysis in medicine. We consider another meta-analysis problem in the context of a decision problem in Section 9.2.

The data in our medical example are displayed in the first three columns of Table 5.4, which summarize mortality after myocardial infarction in 22 clinical trials, each consisting of two groups of heart attack patients randomly allocated to receive or not receive beta-blockers (a family of drugs that affect the central nervous system and can relax the heart muscles). Mortality varies from 3% to 21% across the studies, most of which show a modest, though not 'statistically significant,' benefit from the use of beta-blockers. The aim of a meta-analysis is to provide a combined analysis of the studies that indicates the overall strength of the evidence for a beneficial effect of the treatment under study. Before proceeding to a formal meta-analysis, it is important to apply rigorous criteria in determining which studies are included. (This relates to concerns of ignorability in data collection for observational studies, as discussed in Chapter 8.)