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

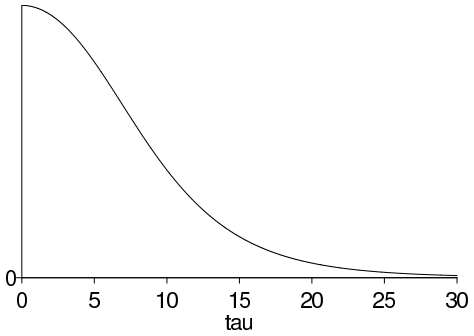

Figure 5.5 *Marginal posterior density, $p(\tau|y)$, for standard deviation of the population of school effects θ_j in the educational testing example.*

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

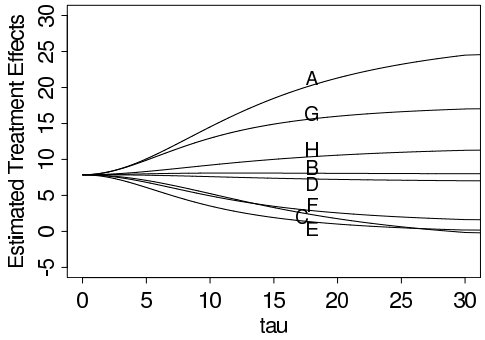

Figure 5.6 *Conditional posterior means of treatment effects,* $E(\theta_j|\tau, y)$*, as functions of the between-school standard deviation* τ*, for the educational testing example. The line for school* C *crosses the lines for* E *and* F *because* C *has a higher measurement error (see Table 5.2) and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

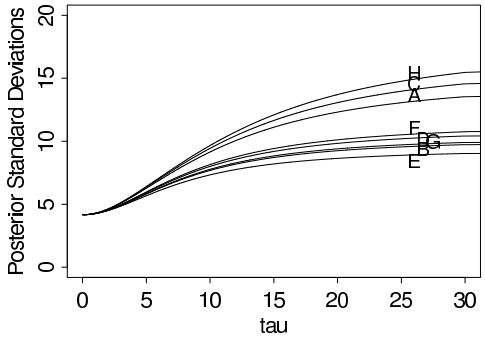

Figure 5.7 *Conditional posterior standard deviations of treatment effects,* $\mathrm{sd}(\theta_j|\tau, y)$*, as functions of the between-school standard deviation* τ*, for the educational testing example.*

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.*

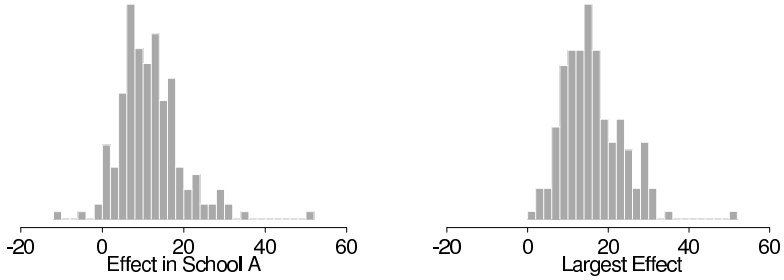

Figure 5.8 *Histograms of two quantities of interest computed from the 200 simulation draws: (a) the effect in school A,* θ_1*; (b) the largest effect,* $\max\{\theta_j\}$*. The jaggedness of the histograms is just an artifact caused by sampling variability from using only 200 random draws.*

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.)