Bayesian Hierarchical Model

Thomas Ng

School	Estimated treatment effect (y_j)	Standard Error of effect estimate, (σ_j)
A	28.39	14.9
B	7.94	10.2
C	-2.75	16.3
D	6.82	11
E	-0.64	9.4
F	0.63	11.4
G	18.01	10.4
H	12.16	17.6

Table I: Data

I. BACKGROUND

This project reproduces the example in Andrew Gelman's Bayesian Data Analysis, namely "Combining information from educational testing experiments in eight schools".

The data are presented in Table I.

II. BAYESIAN HIERARCHICAL MODEL

Suppose $y_j \sim N(\theta_j, \sigma_j^2)$ for $j = 1, \cdots, J$ with $\theta_j \sim N(\mu, \tau^2)$. Let $\theta = (\theta_1, \cdots, \theta_J)$ and $y = (y_1, \cdots, y_J)$ and assume that $\sigma_1^2, \cdots, \sigma_J^2$ are known. And assume $p(\mu, \tau) \propto 1$.
The Hierarchical Structure:

- Level 1: $y_j \sim N(\theta_j, \sigma_j^2) \quad j = 1, \cdots, J$
- Level 2: $\theta_j \sim N(\mu, \tau^2)$
- Level 3: $p(\mu, \tau) \propto 1$

The joint posterior distribution of the parameters given data can be derived in 2 ways: (i) Factorization, (ii) Bayes Theorem

- Factorization:

$$p(\theta, \mu, \tau^2|y) = \overbrace{p(\theta|\mu, \tau^2, y)}^{(1)}\overbrace{p(\mu|\tau^2, y)}^{(2)}\overbrace{p(\tau^2|y)}^{(3)}$$

- Bayes Theorem (conditional independence between y and $\{\mu, \tau^2\}$ given θ):

$$p(\theta, \mu, \tau^2|y) = \frac{p(y|\theta, \mu, \tau^2)p(\theta, \mu, \tau^2)}{p(y)} \propto \begin{array}{l} p(y|\theta, \mu, \tau^2)p(\theta|\mu, \tau^2)p(\mu, \tau^2) \\ p(y|\theta)p(\theta|\mu, \tau^2)p(\mu, \tau^2) \end{array}$$

A. Mathematical derivation

1. $p(\theta_j|\mu, \tau^2, y)$

$$\begin{aligned} p(\theta_j|\mu, \tau^2, y) \propto\ & p(\theta_j, \mu, \tau^2|y) \\ &= p(y|\theta_j)p(\theta_j|\mu, \tau^2) \\ &= N(\theta_j, \sigma_j^2) \times N(\mu, \tau^2) \quad \text{(normal densities)} \\ &\stackrel{d}{=} N\left(\frac{\tau^2 y_j + \mu\sigma_j^2}{\tau^2 + \sigma_j^2}, \frac{\sigma_j^2\tau^2}{\tau^2 + \sigma_j^2}\right) \end{aligned}$$

Since θ_j are independent and $\theta_j|\mu, \tau^2, y \sim N\left(\frac{\tau^2 y_j + \mu\sigma_j^2}{\tau^2 + \sigma_j^2}, \frac{\sigma_j^2\tau^2}{\tau^2 + \sigma_j^2}\right)$, therefore

$$\theta|\mu, \tau^2, y \sim \prod_{j=1}^{J} N\left(\frac{\tau^2 y_j + \mu\sigma_j^2}{\tau^2 + \sigma_j^2}, \frac{\sigma_j^2\tau^2}{\tau^2 + \sigma_j^2}\right)$$

2. $p(\mu|\tau^2, y)$

(Level 1 and Level 3 relationship)

- we avoid some messy algebra by noting that

$$y_j = \theta_j \mathbb{1} + e_j \quad \text{with } \theta_j \sim N(\mu, \tau^2) \text{ and } e_j \sim N_{n_j}(0, \sigma^2 I)$$

, where $\mathbb{1}$ is a $(n_j \times 1)$vector of 1's and θ_j and e_j are independent. Therefore,

$$y_j|\mu, \tau^2, \sigma^2 \sim N_{n_j}\left(\mu, \tau^2 \mathbb{1}\mathbb{1}^T + \sigma^2 I\right)$$

- Consider the joint posterior of $\{\mu, \tau\}$ and by Bayes Theorem and proportionality:

$$p(\mu, \tau^2, y) \propto p(\mu, \tau^2) p(y|\mu, \tau)$$

$$\begin{aligned} p(\mu|\tau^2, y) \propto p(\mu, \tau^2|y) \propto \; & p(\mu, \tau) \prod_{j=1}^{J} p(y_j|\mu, \tau^2) \\ \stackrel{d}{=} & N_{n_j}\left(\mu, \tau^2 \mathbb{1}\mathbb{1}^T + \sigma^2 I\right) \end{aligned}$$

So,

$$\mu|\tau^2, y \sim N\left(\frac{\sum_j \frac{y_j}{\tau^2+\sigma_j^2}}{\sum \frac{1}{\tau^2+\sigma_j^2}}, \frac{1}{\sum_j \frac{1}{\tau^2+\sigma_j^2}}\right)$$

3. $p(\tau|y)$

- The marginal posterior of τ **(independent of μ)** :

$$\begin{aligned} p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)} = & \frac{p(\mu, \tau) \prod_j p(y_j|\mu, \tau^2, \sigma_j^2)}{p(\mu|\tau, y)} \\ \stackrel{d}{=} & \frac{\prod_j N_{n_j}\left(\mu, \tau^2 \mathbb{1}\mathbb{1}^T + \sigma^2 I\right)}{N\left(\frac{\sum_j \frac{y_j}{\tau^2+\sigma_j^2}}{\sum \frac{1}{\tau^2+\sigma_j^2}}, \frac{1}{\sum_j \frac{1}{\tau^2+\sigma_j^2}}\right)} \\ = & \left(\sum_j \frac{1}{\tau^2+\sigma_j^2}\right)^{-1/2} \prod_{j=1} \left(\sigma_j^2+\tau^2\right)^{-1/2} \\ & \times exp\left(-\frac{1}{2}\left(\sum_j \frac{y_j^2}{\tau^2+\sigma_j^2} - \frac{\left(\sum_j \frac{y_j}{\tau^2+\sigma_j^2}\right)^2}{\sum_j \frac{1}{\tau^2+\sigma_j^2}}\right)\right) \end{aligned}$$

B. Reproduce Computation of the Educational Testing Example

First, we plot the *marginal posterior density* $p(\tau|y)$ in Figure II.1.

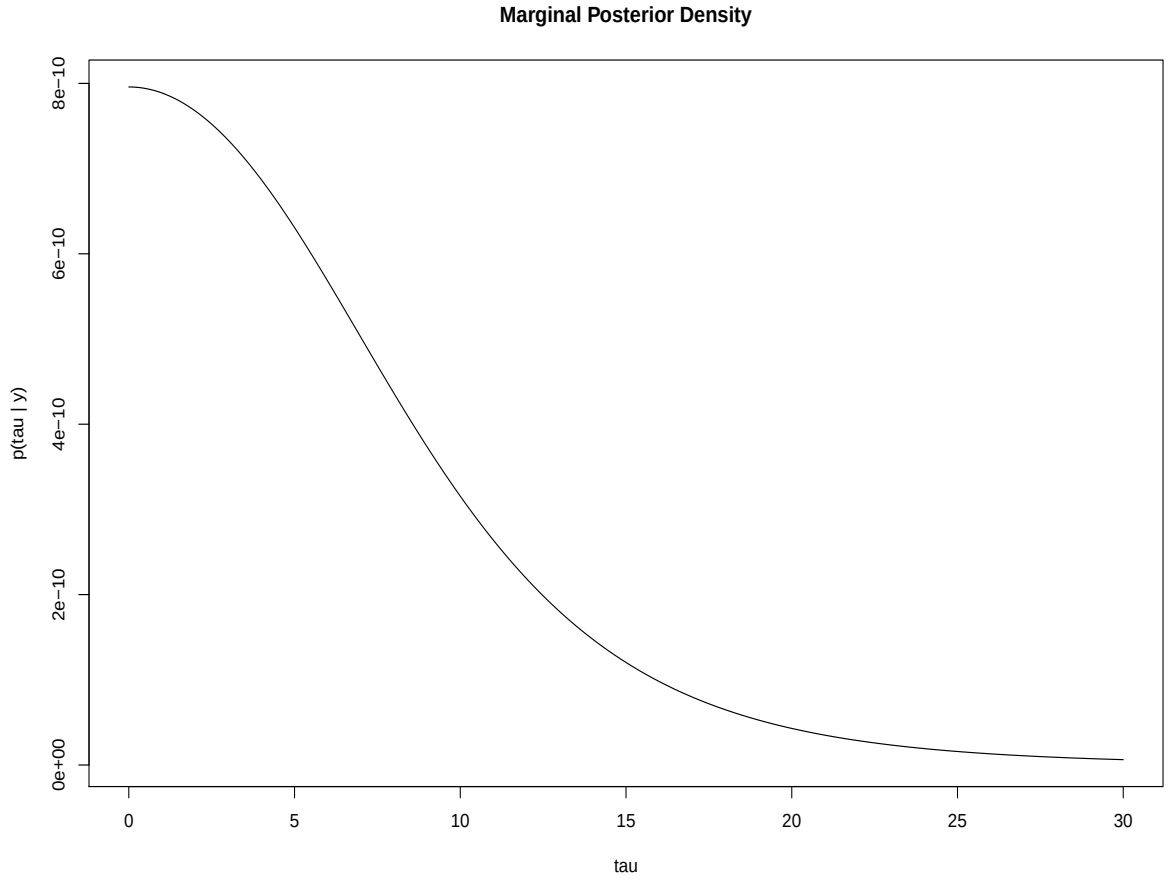

Figure II.1: Marginal Posterior Density $p(\tau|y)$

Second, we plot the *Conditional posterior means of treatment effects* $E(\theta_j|\tau, y)$. See Figure II.2. Note that the line for school C crosses the lines for E and F because C has a higher measurement error. This can be seen in the data table and its estimate is hence shrunk more strongly towards the overall mean.

- By law of iterated expectation,

$$\begin{aligned} E[\theta_j|\tau,y] =& E[E[\theta_j|\mu,\tau,y]|\tau,y] \\ =& E\left[\frac{\tau^2 y_j + \mu\sigma_j^2}{\tau^2+\sigma_j^2}\,\middle|\,\tau,y\right] \\ =& \frac{\tau^2 y_j + \sigma_j^2 E\left[\mu|\tau,y\right]}{\tau^2+\sigma_j^2} \end{aligned}$$

$$= \left[\tau^2 y_j + \sigma_j^2\left(\frac{\sum_j \frac{y_j}{\tau^2+\sigma_j^2}}{\sum_j \frac{1}{\tau^2+\sigma_j^2}}\right)\right]\Big/_{\tau^2+\sigma_j^2}$$

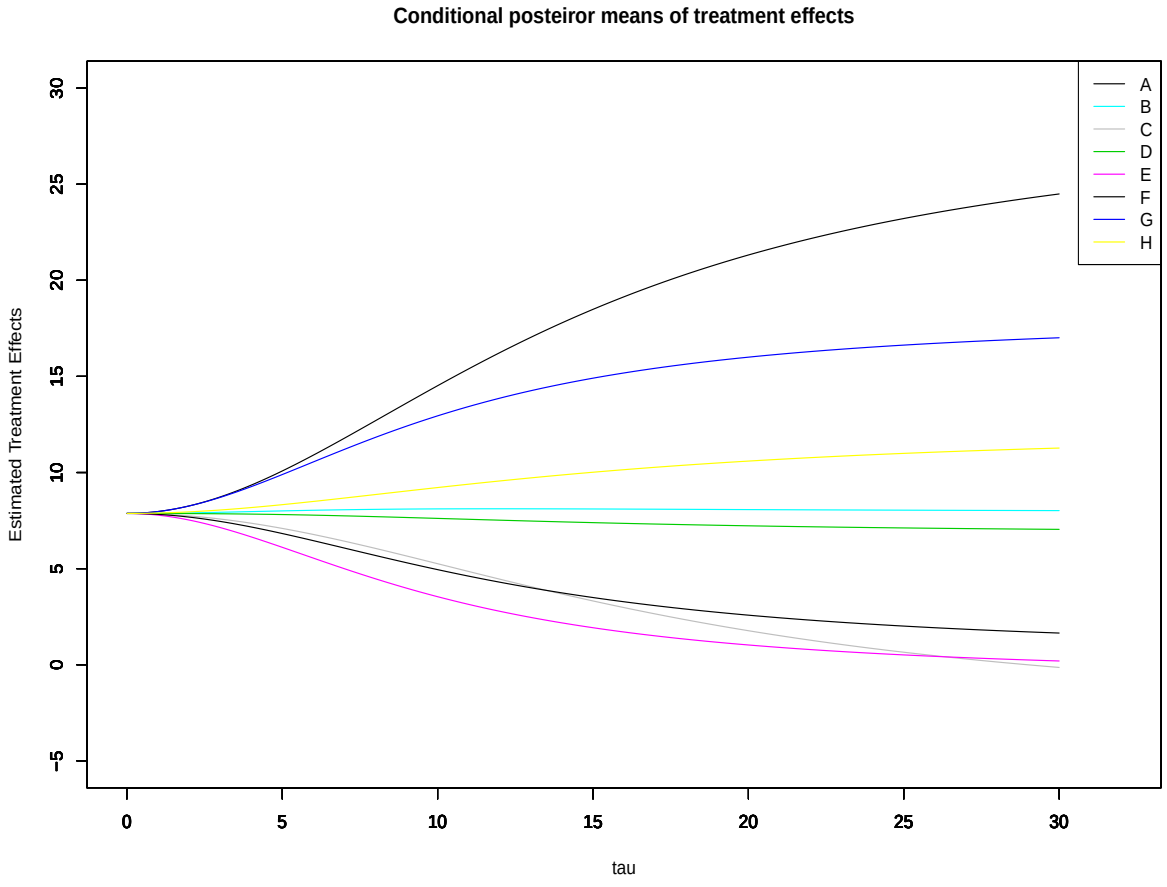

Figure II.2: Conditional posterior means of treatment effects $E(\theta_j|\tau,y)$

Third, we plot the *Conditional posterior standard deviations of treatment effects,* $sd(\theta_j|\tau,y)$. See Figure II.3

$$
\begin{aligned}
Var(\theta_j|\tau,y) =& E\left[\ \mathrm{Var}(\theta_j|\mu,\tau,y)\ \ |\tau,y\right] + \mathrm{Var}\left[E(\theta_j|\mu,j,y)\ \ |\tau,y\right] \\
=& \frac{\sigma_j^2\tau^2}{\tau^2+\sigma_j^2} + \mathrm{Var}\left(\left.\frac{\tau^2 y_j + \mu\sigma_j^2}{\tau^2+\sigma_j^2}\right|\tau,y\right) \\
=& \frac{\sigma_j^2\tau^2}{\tau^2+\sigma_j^2} + \left(\frac{\sigma_j^2}{\tau^2+\sigma^2}\right)^2 \mathrm{Var}(\mu|\tau,y) \\
\\
=& \frac{\sigma_j^2\tau^2}{\tau^2+\sigma_j^2} + \left(\frac{\sigma_j^2}{\tau^2+\sigma^2}\right)^2 \left[\frac{1}{\sum_j {}^1\!/_{\tau^2+\sigma_j^2}}\right].
\end{aligned}
$$

We obtain the posterior standard deviation by taking the square root of above.

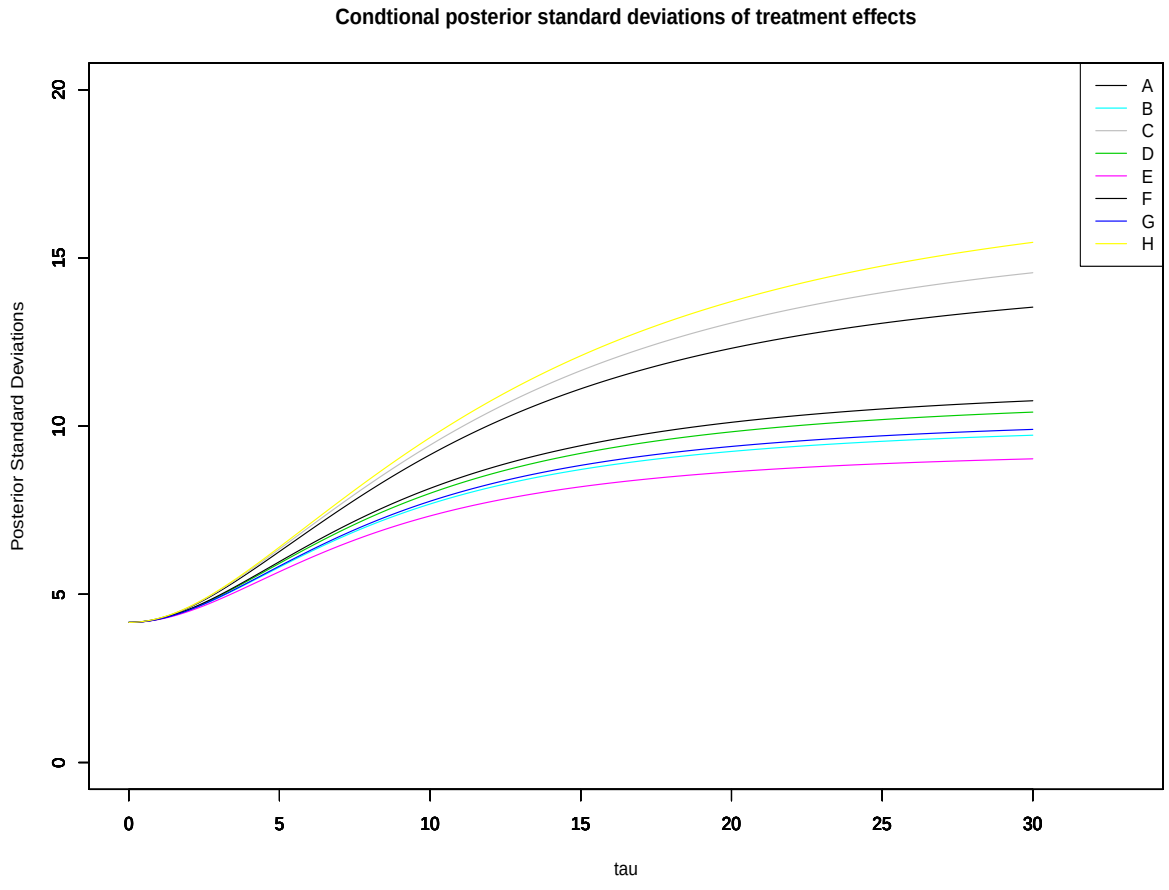

Figure II.3: Conditional posterior standard deviations of treatment effects, $sd(\theta_j|\tau,y)$

Now, we produce the summary table of 200 simulations of treatment effects in the eight schools, we adopt the following theoretical algorithm:

- Sample $\tau^{(t)} \sim p(\tau|y)$ (using the grid method normally, but not in this case).
- Sample $\mu^{(t)} \sim p(\mu|(\tau^2)^{(t)}, y)$
- Sample $\theta_j^{(t)} \sim p(\theta_j|\mu^{(t)}, (\tau^2)^{(t)}, y)$ independently for $j = 1, \cdots, J$

After we sample the θ_j's , we can compute their posterior quantiles (shown below). We make 200 simulation draws.

School Posterior quantiles

	2.5%	25%	50%	75%	97.5%
A	-3	5	10	15	28
B	-5	2	6	12	23
C	-11	-1	5	10	18
D	-7	2	7	12	23
E	-9	-1	3	8	16
F	-9	-1	5	10	19
G	-3	6	11	15	25
H	-9	1	7	12	25

Then, two histograms of two quantities of interested are computed from the 200 simulation draws: (a) the effect in school A, θ_1; (b) the largest effect, $max\{\theta_j\}$. See Figure II.4.

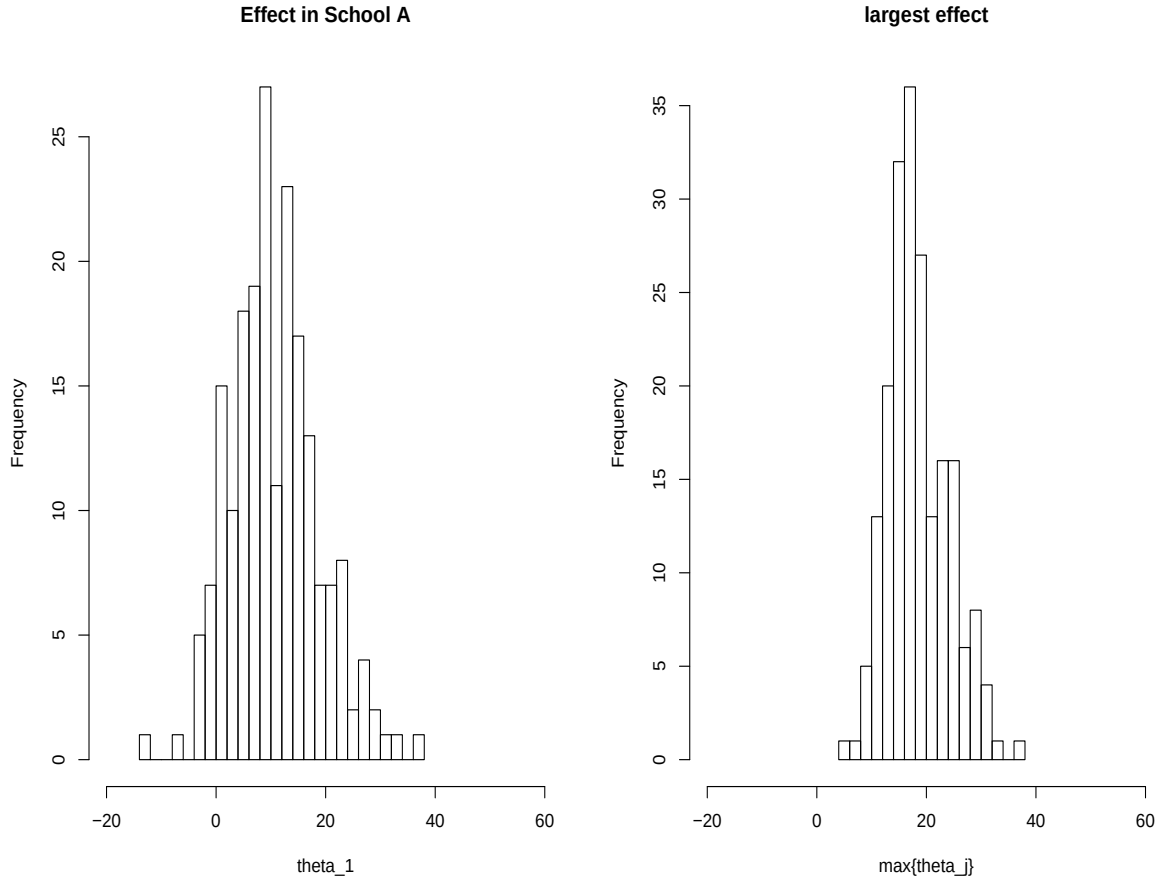

Figure II.4: Histograms: (a) the effect in school A, θ_1; (b) the largest effect, $max\{\theta_j\}$

The histogram on the right depicts the effect of the most successful of the eight coaching programs. It displays a histogram of 200 values from the posterior distribution $p(max\{\theta_j\} > 28..4) \approx {}^{22}/_{200}$. Since the LHS gives the marginal posterior distribution of the effect in school A, and RHS gives the marginal posterior distribution of the largest effect no matter which school it is in, the latter figure has larger values.

C. Part (d) - τ^2

Suppose we want to compute: for each coaching program, the probability that its coaching program is the best. i.e. For School A, we traverse the vector of θ_A s and count the number

School	Probability of this school being the best coaching program
A	0.265
B	0.12
C	0.065
D	0.08
E	0.05
F	0.07
G	0.235
H	0.115

of elements in the vector θ_A that is equal to the $max\{\theta_j\}$. Thus,

$$Pr(\theta_i = max\{\theta_j\}) = Pr(\text{school i is the best}) = \frac{\#\text{ of } \{\theta_i = max\{\theta_j\}\}}{\#\text{ of simulations}}$$

For each pair of schools , j and k, we compute the probability that the coaching program for school j is better than that for school k again by traversing the vectors of θ_j and θ_k and then compare them elements by elements to count how many elements of θ_j is greater than $\theta_k \ \ \forall\ k \neq j$. The matrix below with row j and columns k illustrates this probability. For instance, the first row shows $Pr(\theta_1 > \theta_k) \ \ \forall \ \ k \neq 1$

$$Pr(\theta_j > \theta_k) =$$

$$\begin{bmatrix} - & 0.63 & 0.7 & 0.665 & 0.785 & 0.735 & 0.49 & 0.65 \\ 0.37 & - & 0.595 & 0.48 & 0.66 & 0.555 & 0.35 & 0.465 \\ 0.3 & 0.405 & - & 0.415 & 0.585 & 0.475 & 0.28 & 0.42 \\ 0.335 & 0.52 & 0.585 & - & 0.675 & 0.57 & 0.345 & 0.495 \\ 0.215 & 0.34 & 0.415 & 0.325 & - & 0.435 & 0.22 & 0.415 \\ 0.265 & 0.445 & 0.525 & 0.43 & 0.565 & - & 0.255 & 0.38 \\ 0.51 & 0.65 & 0.72 & 0.655 & 0.78 & 0.745 & - & 0.655 \\ 0.35 & 0.535 & 0.58 & 0.505 & 0.585 & 0.62 & 0.345 & - \end{bmatrix}$$

D. Part (e) - Suppose $\tau^2 = \infty$

When $\tau^2 = \infty$, we can directly sample θ_j from $\theta_j|\mu, \tau^2, y \sim N(y_j, \sigma_j^2)$, since the distribution of θ_j doesnt depend on μ anymore. This means different schools / coaching programs have different means y_j ,i.e. group means have no pattern and nothing in common.

$$E[\theta_j|y, \mu, \sigma^2, \tau^2] = y_j$$

Therefore, individual coaching programs have their own individual means.
Under this assumption, if we repeat the computation of the probabilities in part(d), we obtain:

$$Pr(\theta_i = max\{\theta_j\}) = Pr(\text{school i is the best}) = \frac{\#\text{ of } \{\theta_i = max\{\theta_j\}\}}{\#\text{ of simulations}}$$

(Output can be reproduced if user runs the R code.)

$$Pr(\theta_j > \theta_k) =$$

$$\begin{bmatrix} - & 0.84 & 0.91 & 0.875 & 0.935 & 0.905 & 0.715 & 0.81 \\ 0.16 & - & 0.725 & 0.54 & 0.775 & 0.66 & 0.27 & 0.47 \\ 0.09 & 0.275 & - & 0.33 & 0.5 & 0.38 & 0.14 & 0.32 \\ 0.125 & 0.46 & 0.67 & - & 0.745 & 0.62 & 0.19 & 0.45 \\ 0.065 & 0.225 & 0.5 & 0.255 & - & 0.385 & 0.08 & 0.24 \\ 0.095 & 0.34 & 0.62 & 0.38 & 0.615 & - & 0.14 & 0.295 \\ 0.285 & 0.73 & 0.86 & 0.81 & 0.92 & 0.86 & - & 0.665 \\ 0.19 & 0.53 & 0.68 & 0.55 & 0.76 & 0.705 & 0.335 & - \end{bmatrix}$$

E. Part (f) - Suppose $\tau^2 = 0$

In this case, we have

$$\mu|\tau^2, y \sim N\left(\frac{\sum_j \frac{y_j}{\sigma_j^2}}{\sum_j \frac{1}{\sigma_j^2}}, \frac{1}{\sum_j \frac{1}{\sigma_j^2}}\right)$$

and

$$\theta_j|\mu, \tau^2, y \sim N(\mu, 0)$$

This essentially means $\theta_j = \mu$. Thus, we sample μ from $\mu|\tau^2, y$ and then set $\theta_j = \mu$.
The implication here is that individual coaching programs have the same mean, i.e. group means are all equal. So there is no difference between groups. Thus.

$$E[\theta_j|Y, \mu, \tau^2, \sigma^2] = \mu$$

We shall expect the probability of school j being the best coaching program should be the same across all schools. We shoud also expect the probability that the coaching program for school j is better than that for school k to be roughly the same (i.e. equally likely) for all j, because individual schools have same means, i.e. no difference between them.
Thus, we obtain:

$$Pr(\theta_i = max\{\theta_j\}) = Pr(\text{school i is the best}) = \frac{\#\text{ of } \{\theta_i = max\{\theta_j\}\}}{\#\text{ of simulations}}$$

(Output can be reproduced if user runs the R code.)

$$Pr(\theta_j > \theta_k) =$$

$$\begin{bmatrix}
- & 0.515 & 0.46 & 0.46 & 0.515 & 0.49 & 0.52 & 0.525 \\
0.485 & - & 0.48 & 0.47 & 0.455 & 0.525 & 0.495 & 0.515 \\
0.54 & 0.52 & - & 0.5 & 0.485 & 0.51 & 0.505 & 0.5 \\
0.54 & 0.53 & 0.5 & - & 0.51 & 0.54 & 0.49 & 0.57 \\
0.485 & 0.545 & 0.515 & 0.49 & - & 0.485 & 0.495 & 0.51 \\
0.51 & 0.475 & 0.49 & 0.46 & 0.515 & - & 0.47 & 0.515 \\
0.48 & 0.505 & 0.495 & 0.51 & 0.505 & 0.53 & - & 0.55 \\
0.475 & 0.485 & 0.5 & 0.43 & 0.49 & 0.485 & 0.45 & -
\end{bmatrix}$$

F. Discussion: How do the results differ with different values of τ^2

In part (d), we have the bayesian continuum. Part (e) and (f) represent the limit case, where we get the classical dichotomy, while part d repsents the bayesian continuum , i.e. a

weighted average of the 2 extremes.

In part (e), when $\tau^2 \to \infty$, this means group means have no pattern, nothing in common. The estimated treatment effect is essentially the observed treatment effect: $\hat{\theta}_j = y_j$. This basically reiterates the results of data and hence, uninformative, as we can simply observe from the observed treatment effect that school A has the best coaching program.

In part (f), when $\tau^2 \to 0$, this means group means are all equal, i.e. $\hat{\theta}_j = E[\mu|y, \sigma^2, \tau^2] = \bar{y}$... In this case, all coaching programs are equally as good. Each school can be the best coaching program with probability $\frac{1}{8}$. This is uninformative either, since setting $\tau^2 = 0$ assume there is no difference between groups and thus obscures what the data is trying to tell us.

Relatively speaking, Part (d) is an informative, as it is the weighted average of both extremes in part (e) and (f). It constitutes a comparison with the classical dichotomy through **the Shrinkage Parameter**:

The *Shrinkage Parameter for school(subgroup j)* is:

$$B_j = \frac{1/\tau^2}{1/\sigma_j^2 + 1/\tau^2} = \frac{\sigma_j^2}{\tau^2 + \sigma_j^2}$$

Our estimate of $(\theta_j - \mu)$ is $(y_j - \mu)$ "shrunk by B_j", *i.e.shrink group means towards global means*

$$E[\theta_j|Y, \mu, \sigma^2, \tau^2] - \mu = (1 - B_j)(y_j - \mu)$$

Copying from the book:

To summarise, the Bayesian analysis of this example not only allows straightforward inference about many inferences about many parameters that may be of interest, but the hierarchical model is flexible enough to adapt to the data, thereby providing posterior inferences that account for the **partial pooling** as well as the uncertainty in the hyperparameters.