below.

$$\Pr(\theta_i \text{ is the largest}) = \int_{-\infty}^{\infty} \prod_{j \neq i} \Phi\left(\frac{\theta_i - \theta_j}{\sigma_j}\right) \phi(\theta_i | y_i, \sigma_i) d\theta_i$$

This integral can be evaluated numerically (results given below) or estimated by simulating school effects from their independent normal posterior distributions. The results are provided below.

| | | Pr(better than school) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| School | Pr(best) | A | B | C | D | E | F | G | H |
| A | 0.556 | – | 0.87 | 0.92 | 0.88 | 0.95 | 0.93 | 0.72 | 0.76 |
| B | 0.034 | 0.13 | – | 0.71 | 0.53 | 0.73 | 0.68 | 0.24 | 0.42 |
| C | 0.028 | 0.08 | 0.29 | – | 0.31 | 0.46 | 0.43 | 0.14 | 0.27 |
| D | 0.034 | 0.12 | 0.47 | 0.69 | – | 0.70 | 0.65 | 0.23 | 0.40 |
| E | 0.004 | 0.05 | 0.27 | 0.54 | 0.30 | – | 0.47 | 0.09 | 0.26 |
| F | 0.013 | 0.07 | 0.32 | 0.57 | 0.35 | 0.53 | – | 0.13 | 0.29 |
| G | 0.170 | 0.28 | 0.76 | 0.86 | 0.77 | 0.91 | 0.87 | – | 0.61 |
| H | 0.162 | 0.24 | 0.58 | 0.73 | 0.60 | 0.74 | 0.71 | 0.39 | – |

**5.1c.** The model with $\tau$ set to $\infty$ has more extreme probabilities. For example, in the first column, the probability that School A is the best increases from 0.25 to 0.56. It is also true in the pairwise comparisons. For example, the probability that School A's program is better than School E is under the full hierarchical model is 0.73, whereas it is 0.95 under the $\tau = \infty$ model. The more conservative answer under the full hierarchical model reflects the evidence in the data that the coaching programs appear fairly similar in effectiveness. Also, noteworthy is that the preferred school in a pair can change, so that School E is better than School C when $\tau$ is set to $\infty$, whereas School C is better than School E when averaging over the posterior distribution of $\tau$. This effect occurs only because the standard errors, $\sigma_j$, differ.

**5.1d.** If $\tau = 0$ then all of the school effects are the same. Thus no school is better or worse than any other.

**5.2a.** Yes, they are exchangeable. The joint distribution is

$$p(\theta_1, \ldots, \theta_{2J}) = \binom{2J}{J}^{-1} \sum_p \left( \prod_{j=1}^{J} \mathrm{N}(\theta_{p(j)} | 1, 1) \prod_{j=J+1}^{2J} \mathrm{N}(\theta_{p(j)} | -1, 1) \right), \tag{6}$$

where the sum is over all permutations $p$ of $(1, \ldots, 2J)$. The density (6) is obviously invariant to permutations of the indexes $(1, \ldots, 2J)$.

**5.2b.** Pick any $i, j$. The covariance of $\theta_i, \theta_j$ is *negative*. You can see this because if $\theta_i$ is large, then it probably comes from the $\mathrm{N}(1, 1)$ distribution, which means that it is more likely than not that $\theta_j$ comes from the $\mathrm{N}(-1, 1)$ distribution (because we know that exactly half of the $2J$ parameters come from each of the two distributions), which means that $\theta_j$ will probably be negative. Conversely, if $\theta_i$ is negative, then $\theta_j$ is most likely positive.

Then, by Exercise 5.3, $p(\theta_1, \ldots, \theta_{2J})$ cannot be written as a mixture of iid components.

The above argument can be made formal and rigorous by defining $\phi_1, \ldots, \phi_{2J}$, where half of the $\phi_j$'s are 1 and half are $-1$, and then setting $\theta_j | \phi_j \sim \mathrm{N}(\phi_j, 1)$. It's easy to show first that $\mathrm{cov}(\phi_i, \phi_j) < 0$, and then that $\mathrm{cov}(\theta_i, \theta_j) < 0$ also.

**5.2c.** In the limit as $J \to \infty$, the negative correlation between $\theta_i$ and $\theta_j$ approaches zero, and the joint distribution approaches iid. To put it another way, as $J \to \infty$, the distinction disappears between (1) independently assigning each $\theta_j$ to one of two groups, and (2) picking exactly half of the $\theta_j$'s for each group.

**5.3.** Let $\mu(\phi) = \mathrm{E}(\theta_j|\phi)$. From (1.8) on page 24 (also see Exercise 1.2),

$$\begin{aligned}\mathrm{cov}(\theta_i, \theta_j) &= \mathrm{E}(\mathrm{cov}(\theta_i, \theta_j|\phi)) + \mathrm{cov}(\mathrm{E}(\theta_i|\phi), \mathrm{E}(\theta_j|\phi)) \\ &= 0 + \mathrm{cov}(\mu(\phi), \mu(\phi)) \\ &= \mathrm{var}(\mu(\phi)) \\ &\geq 0.\end{aligned}$$

**5.5a.** We want to find $\mathrm{E}(y)$ and $\mathrm{var}(y)$, where $y|\theta \sim \mathrm{Poisson}(\theta)$ and $\theta \sim \mathrm{Gamma}(\alpha, \beta)$. From (2.7), $\mathrm{E}(y) = \mathrm{E}(\mathrm{E}(y|\theta))$; from properties of the Poisson and Gamma distributions, we have $\mathrm{E}(y|\theta) = \theta$ so that $\mathrm{E}(\mathrm{E}(y|\theta)) = \mathrm{E}(\theta) = \alpha/\beta$.

Similarly, by formula (2.8),

$$\mathrm{var}(y) = \mathrm{E}(\mathrm{var}(y|\theta)) + \mathrm{var}(\mathrm{E}(y|\theta)) = \mathrm{E}(\theta) + \mathrm{var}(\theta) = \alpha/\beta + \alpha/\beta^2 = \alpha/\beta^2(\beta + 1).$$

**5.5b.** This part is a little bit trickier because we have to think about what to condition on in applying formulas (2.7) and (2.8). Some reflection (and, perhaps, a glance at formula (3.3) on page 75) will lead to the choice of conditioning on $\sigma^2$.

Throughout these computations, $n$, $s$, and $\bar{y}$ are essentially treated like constants.

From (2.7) and (3.3),

$$\mathrm{E}(\sqrt{n}(\mu - \bar{y})/s|y) = \mathrm{E}(\mathrm{E}(\sqrt{n}(\mu - \bar{y})/s|\sigma, y)|y) = \mathrm{E}((\sqrt{n}/s)\mathrm{E}(\mu - \bar{y}|\sigma, y)|y) = \mathrm{E}((\sqrt{n}/s) \cdot 0|y) = \mathrm{E}(0|y) = 0.$$

Obviously we must have $n > 1$ for $s$ to be defined. But in order for the expectation to exist, we must have $n > 2$. You can deduce this from inspecting the formula for $p(\mu|y)$ at the top of page 69: the exponent must be less than negative one for the quantity to be integrable, which is true if and only if $n > 2$.

Similarly, we can compute from (2.8), (3.3), and (3.5) that

$$\begin{aligned}\mathrm{var}(\sqrt{n}(\mu - \bar{y})/s|y) &= \mathrm{var}(\mathrm{E}(\sqrt{n}(\mu - \bar{y})/s|\sigma, y)|y) + \mathrm{E}(\mathrm{var}(\sqrt{n}(\mu - \bar{y})/s|\sigma, y)|y) \\ &= \mathrm{var}(0|y) + \mathrm{E}((n/s^2)\mathrm{var}(\mu|\sigma, y)|y) \\ &= \mathrm{E}((n/s^2)\sigma^2/n|y) \\ &= \mathrm{E}(\sigma^2|y)/s^2 = \frac{n-1}{n-3}.\end{aligned}$$

Note that $n > 3$ is required here.

**5.6.** Let $p_m(\theta|y)$ denote the posterior density of $\theta$ corresponding to the prior density $p_m(\theta)$. That is, for each $m$, $p_m(\theta|y) = p_m(\theta)p(y|\theta)/p_m(y)$, where $p_m(y)$ is the prior predictive density.

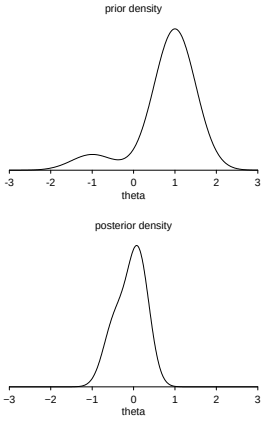
If $p(\theta) = \sum_m \lambda_m p_m(\theta)$, then the posterior density of $\theta$ is proportional to $\sum_m \lambda_m p_m(\theta)p(y|\theta) = \sum_m \lambda_m p_m(y)p_m(\theta|y)$: this is a mixture of the posterior densities $p_m(\theta|y)$ with weights proportional to $\lambda_m p_m(y)$. Since each $p_m(\theta)$ is conjugate for the model for $y$ given $\theta$, the preceding computation demonstrates that the class of finite mixture prior densities is also conjugate.

Consider an example: $p_1(\theta) \sim \mathrm{N}(1, 0.5^2)$, $p_2(\theta) \sim \mathrm{N}(-1, 0.5^2)$, and suppose that $\lambda_1 = 0.9$ amd $\lambda_2 = 0.1$. (The exact choices of $\lambda_1$ and $\lambda_2$ are not important. What is important is how and why the posterior mixture weights are different than the prior weights.) We know that $p_1(\theta|y) \sim$

$\mathrm{N}(1.5/14, 1/14)$ and that $p_2(\theta|y) \sim \mathrm{N}(-6.5/14, 1/14)$; see, e.g., formulas (2.9) and (2.10). We also know from the last paragraph that $p(\theta|y)$ will be a weighted sum of these conditional posterior densities with weights $\lambda_m p_m(y)/\sum_k \lambda_k p_k(y)$ for $m = 1, 2$.

You can compute $p_1(y)$ and $p_2(y)$, using convenient properties of normal distributions: $p_1(y) = \mathrm{N}(-0.25|1, 0.5^2 + 1/10) = 0.072$, and $p_2(y) = \mathrm{N}(-0.25| - 1, 0.5^2 + 1/10) = 0.302$.

So the weights for $p_1(\theta|y)$ and $p_2(\theta|y)$ are not 0.9 and 0.1 but are, rather, $\frac{0.9 \cdot 0.072}{0.9 \cdot 0.072 + 0.1 \cdot 0.302} = 0.68$ and $\frac{0.1 \cdot 0.302}{0.9 \cdot 0.072 + 0.1 \cdot 0.302} = 0.32$.




```
theta <- seq(-3,3,.01)
prior <- c (0.9, 0.1)
dens <- prior[1]*dnorm(theta,1,0.5) +
  prior[2]*dnorm(theta,-1,0.5)
plot (theta, dens, ylim=c(0,1.1*max(dens)),
  type="l", xlab="theta", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
mtext ("prior density", cex=2, 3)

marg <- dnorm(-.25,c(1,-1),sqrt(c(0.5,0.5)^2+1/10))
posterior <- prior*marg/sum(prior*marg)

dens <- posterior[1]*dnorm(theta,1.5/14,sqrt(1/14)) +
  posterior[2]*dnorm(theta,-6.5/14,sqrt(1/14))
plot (theta, dens, ylim=c(0,1.1*max(dens)),
  type="l", xlab="theta", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
mtext ("posterior density", cex=2, 3)
```

**5.7a.** Consider the limit $(\alpha + \beta) \to \infty$ with $\alpha/\beta$ fixed at any nonzero value. The likelihood (see equation (5.8)) is

$$\begin{aligned} p(y|\alpha, \beta) &\propto \prod_{j=1}^{J} \frac{[\alpha \cdots (\alpha + y_j - 1)][\beta \cdots (\beta + n_j - y_j - 1)]}{(\alpha + \beta) \cdots (\alpha + \beta + n_j - 1)} \qquad (7) \\ &\approx \prod_{j=1}^{J} \frac{\alpha^{y_j} \beta^{n_j - y_j}}{(\alpha + \beta)^{n_j}} \\ &= \prod_{j=1}^{J} \left(\frac{\alpha}{\alpha + \beta}\right)^{y_j} \left(\frac{\beta}{\alpha + \beta}\right)^{n_j - y_j}, \end{aligned}$$

which is a constant (if we are considering $y$, $n$, and $\alpha/\beta$ to be fixed), so the prior density determines whether the posterior density has a finite integral in this limit. A uniform prior density on $\log(\alpha+\beta)$ has an infinite integral in this limit, and so the posterior density does also in this case.

**5.7b.** The Jacobian of the transformation is

$$\left| \begin{array}{cc} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ -\frac{1}{2}(\alpha+\beta)^{-3/2} & -\frac{1}{2}(\alpha+\beta)^{-3/2} \end{array} \right| = \text{constant} \cdot (\alpha + \beta)^{-5/2}.$$

**5.7c.** Note error in statement of problem in the first and second printings: "improper" should be "proper" and "$y_j \neq 0$ for at least one experiment $j$ and $y_j \neq n_j$ for at least one experiment $j$" should be "$0 < y_j < n_j$ for at least one experiment $j$."

There are 4 limits to consider: