7.1 Exercises

1. In the schools example, (see lecture notes), what is the posterior probability that the coaching effect in school A was higher than in school G?

2. Write the WinBUGS code and simulate results from a hierarchical normal model where:

$$\begin{aligned} y_{ij} &\sim \text{Normal}(\theta_i, 1) \\ \theta_i &\sim \text{Normal}(\mu_\theta, \sigma_\theta^2) \\ \sigma_\theta &\sim \text{Uniform}(0, 1000) \\ \mu_\theta &\sim \text{Normal}(0, \text{large } \sigma^2) \end{aligned}$$

```
list(y=structure(.Data=c(
  6.30,    5.99,    4.85,    3.60,    5.71,
  0.82,    2.45,    2.34,    3.52,    2.42,
  8.60,    9.63,    8.24,    9.23,    9.20,
  2.58,    2.27,    3.68,    0.71,    3.01),.Dim=c(4,5)))
```

Try replacing the prior of σ_θ by constant values in the range $\{0.01, \ldots, 1000\}$. Observe what happens to the posterior means and variances of θ_i? Replace the unit variance of y_{ij} by σ_y^2 and assign a prior for this, and repeat computations. Compute the posterior of $\sigma_\theta^2/(\sigma_y^2 + \sigma_\theta^2)$. You may also generate your own data with chosen parameters and check how accurate estimates you get.

3. Assume the true disease incidence varies over a number of geographical regions i, e.g. $i = 1, \ldots, 100$. To make your own data, generate these true values λ_i from Gamma$(5, 1000)$. Then, for each generated λ_i, generate the actual disease count X_i for each region from Poisson$(N_i\lambda_i)$, assuming the exposure (population count) is $N_i = i \times 10$. Now, you know the 'true values', and you have 'observed data'. Compute the observed incidences $\hat{\lambda}_i = X_i/N_i$ and plot them as a function of population size. Observe how these behave when N_i is small. Compute the posterior distribution of all λ_i assuming the prior $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, $\alpha \sim \text{Exp}(0.01)$, $\beta \sim \text{Exp}(0.01)$. Compute the posterior medians and 95% CIs and compare them with the observed incidences, $\hat{\lambda}_i$, as a function of population size.

4. 60 egg producing flocks are tested for salmonella. From each flock 4000 eggs are collected and analyzed in pools of 10 eggs. Each flock has its own prevalence p_i of salmonella contaminated eggs. When the pooled sample is analyzed, if any of the 10 eggs are contaminated, the test result will be positive with probability p_s (sensitivity). If none of the 10 eggs are contaminated, the test result will be negative. Assume a hyper prior for p_i to describe the larger population of flocks so that $\text{logit}(p_i) \sim \text{N}(\mu, \tau)$ with $\mu \sim \text{N}(0, 0.001)$ and $\tau = 1/\sigma^2 \sim \text{Gamma}(0.01, 0.01)$. Prior of test sensitivity p_s is a beta-density elicited from expert opinion. According to this, the sensitivity should be approximately between 0.3 and 0.7. (This should account for the uncertainty concerning within flock prevalence affecting the sampling, and also laboratory test sensitivity). Using WinBUGS, compute posterior predictive distribution for egg prevalence in a next flock from this population of flocks.

```
list(X=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 1,1,1,1,1,1,1,1,2,2,3,3,4,4,4,6,8,25))
```

5. Generate 10 random variables from $\mathrm{N}(0, \sigma^2)$ and 30 random variables from $\mathrm{N}(3, \sigma^2)$ with fixed $\sigma = 1$. Based on the full set of data, $X_1, \ldots, X_{40}$, compute posterior distribution assuming the mixture model $X_i \sim \sum_{i=1}^{2} w_i \mathrm{N}(\mu_i, \sigma^2)$, with priors $w_i = \mathrm{U}(0,1)$ and $\mu_i = \mathrm{N}(0, \text{large variance})$. Define latent indicators for group membership and, using them, compute the posterior distribution of the unknown group sizes in the actual data set. Study how the result depends on chosen values of σ. (If σ_i was treated as an unknown parameter in both component densities of the mixture, together with unknown μ_i, some restriction for these variance parameters is needed for identifiability).

6. Make a WinBUGS code and compute the posterior distribution in the example of rat tumors. (Data: Gelman p. 119). $X_i \sim \mathrm{Bin}(N_i, \theta_i)$, $\theta_i \sim \mathrm{Beta}(\alpha, \beta)$. $((\frac{\alpha}{\alpha+\beta}), (\alpha+\beta)^{-1/2}) \propto 1$.

```
list(x=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1, 2,2,2,2,2,2,2,2,2, 1, 5, 2, 5, 3, 2, 7,7, 3,3, 2,
9, 10, 4,4,4,4,4,4,4, 10, 4,4,4, 5,11, 12, 5,5, 6, 5, 6,6,6,6, 16,
15,15, 9), n=c( 20,20,20,20,20,20,20, 19,19,19,19, 18,18, 17,
20,20,20,20, 19,19, 18,18, 25, 24, 23, 20,20,20,20,20,20,
10,49,19,46,27,17,49,47,20,20,13,48,50,20,20,20,20,20,20,20,
48,19,19,19,22,46,49,20,20,23,19,22,20,20,20,52,47,46,24) )
```

7. Make a WinBUGS code and compute the posterior distribution of farm specific prevalence of tonsil positive pig carcasses from the data (15 farms):

```
X=c(14,18,16,20,0,8,9,12,20,14,12,20,21,9,9)
N=c(24,25,18,24,23,25,26,24,24,21,22,25,24,25,20)
```

with the model: $X_i \sim \mathrm{Bin}(N_i, p_i)$, $p_i \sim \mathrm{Beta}(\alpha, \beta)$, $\alpha \sim \mathrm{Gamma}(2, 1/2)$, $\beta \sim \mathrm{Gamma}(2, 1/2)$. Compare the posterior predictive distribution of p with prior predictive distribution. Check how much the prediction is influenced by the 5th observation (quite extreme value) by removing that from data. Sensitivity to possible outliers could be checked like this, but also sensitivity to different priors could be checked.

8. Use the kangaroo skull data from the lectures and compute posterior predictive distribution for female and male skulls. Define a few artificial 'measurements' X ranging from 700 to 2200 and compute the posterior probability of male for each (Note that all fixed values become part of data, from which the posterior is computed). Check the resulting probability for different values of X.