

Name : -----

PRACTICE PROBLEMS FOR MIDTERM 2 in 2004
Bayesian Statistics, 22S:138
Fall 2003, Instructor: Cowles
Midterm 2

Show any computations that you carry out.

1. Reconsider the “Dyes” example from homework 6, in which the observed data were the yields of dyestuff y_{ij} for 5 samples from each of 6 randomly chosen batches of raw material. The subscript $i, i = 1, \dots, 6$, indicates the batch, and the subscript $j, j = 1, \dots, 5$, identifies the sample within the batch. In the WinBUGS example and in homework 6, it was assumed that the within-batch precision τ_{within}^2 of dyestuff yield was the same for all batches, but that each batch i had its own population mean μ_i . This model can be expressed as:

Model 1

Likelihood:

$$y_{ij} | \mu_i, \tau_{within}^2 \sim N(\mu_i, \tau_{within}^2)$$

Second stage:

$$\mu_i | \theta, \tau_{betw}^2 \sim N(\theta, \tau_{betw}^2)$$

Third stage:

$$\begin{aligned} \theta &\sim N(0, 0.000001) \\ \tau_{within}^2 &\sim G(0.001, 0.001) \\ \tau_{betw}^2 &\sim G(20, 20000) \end{aligned}$$

Suppose that instead, we believed that the means were the same for all batches. However, we now relax the assumption of equal within-batch precisions, and allow each batch to have its own precision $\tau_{within,i}^2$.

- (a) Write the following changes into the specification of Model 1.
- Change the likelihood to assume a common mean for all batches but to allow for individual precisions $\tau_{within,i}^2$.
 - To the second stage, add a semi-conjugate prior for the $\tau_{within,i}^2$ s. The parameters of this semi-conjugate prior should be unknown parameters that will be estimated. Delete any items that are no longer needed in the second stage.

- Complete the model specification by changing, adding and/or deleting necessary items in the third stage. If you add any new prior(s) at the third stage, make them vague but proper.
- Using your new model, write the expression to which the joint posterior distribution of all unknown model parameters given the observed data is proportional.

- (c) Derive the posterior full conditional for $\tau_{with,2}^2$. If possible, identify it as a standard probability distribution.

2. Assume the following notation:

$\mathbf{y} = y_1, \dots, y_n$ are the observed data

$\boldsymbol{\theta} = \theta_1, \dots, \theta_p$ are all the unknown parameters in the model

$\mathbf{ynew} = ynew_1, \dots, ynew_n$ are future data that have not been observed yet

We have studied the following types of distributions that arise in Bayesian statistics.

- full conditional distribution
- joint posterior distribution
- marginal posterior distribution
- posterior predictive distribution
- prior distribution

For each of the distributions below, write which type it is (from the above list):

- $p(\mathbf{ynew} | \mathbf{y})$
- $p(\boldsymbol{\theta} | \mathbf{y})$
- $p(\theta_1 | \mathbf{y}, \theta_2, \dots, \theta_p)$
- $p(\theta_1)$
- $p(\theta_1 | \mathbf{y})$

3. NOTE: On Mon. I will give you a paper handout of the WinBUGS graphs and tables that you need to answer some parts of this question. But you can do the first and last sections (on exchangeability and stages of the model) without them. The WinBUGS code and data are on the last page of this paper.

In each of $N = 10$ cities in El Salvador, people were tested in the early 1970's for the disease toxoplasmosis. We have the following variables measured in each city:

$n[i]$ number of people tested in city i , $i = 1, \dots, N$
 $r[i]$ number who tested positive for toxoplasmosis in city i

The primary research question was the proportion of urban residents of the country that were infected with toxoplasmosis.

See the attached WinBUGS code and output concerning a model that was fit to these data. (Graphs for $p[2]$ through $p[10]$ were very similar to those for $p[1]$ and are omitted.)

- (a) From the following list, which quantities are treated as exchangeable in this model? (Circle as many as appropriate.)
- all the individual outcomes (positive/not positive) in all the people tested in all cities
 - all the individual outcomes (positive/not positive) in all the people tested in a single city
 - all the success probabilities p_i in all cities
 - none of the above
- (b) What are the posterior mean and the 95% central posterior interval for the parameter of primary interest (the population proportion of the urban population infected with toxoplasmosis at the time of the study)? (numeric answers)
- (c) How many initial iterations would you discard? (Alpha was the parameter with the worst values of Gelman and Rubin diagnostic.)
 Give a numeric answer, and briefly state how you determined it.

- (d) The autocorrelations between samples of the parameter α taken 10 iterations apart is approximately:
- 1.0
 - 0.5
 - 0.0
 - impossible to determine from the output provided
- (e) The Monte Carlo error for $p[2]$ is .0017. This is (circle one)
- the probability that the MCMC sampler made an error
 - a quantity that will not change regardless of how many iterations are run
 - the estimated standard deviation of observations drawn from the population
 - a measure of the uncertainty in estimating the posterior mean of $p[2]$ by using the sampler output
 - none of the above

(f) Fill in the following table with an entry for each stage of the model:

—Name of stage—	—Line(s) of WinBUGS code that specify it—

WinBUGS code for problem 3

```

model
{
  for (i in 1:N)
  {
    r[i] ~ dbin(p[i], n[i])
    p[i] ~ dbeta( alpha, beta)
  }
  alpha ~ dgamma(0.1, 0.1)
  beta ~ dgamma( 0.1, 0.1)
  P <- alpha / (alpha + beta)
}

Data
list(n=c(51,16,82, 13, 43,75,13,10,6,37), r=c(24,7,46,9,23,53,8,3,1,23),
N = 10)

Inits
list(p = c(.5, .5, .5, .5, .5, .5, .5, .5, .5, .5), alpha = 1, beta = 1)
list(p = c(.1, .1, .1, .1, .1, .1, .1, .1, .1, .1), alpha = .1, beta = .9)
list(p = c(.9, .9, .9, .9, .9, .9, .9, .9, .9, .9), alpha = 9, beta = 1)

```