MAS3301 Bayesian Statistics
Problems 4 and Solutions

Semester 2

2008-9

Problems 4

1. I recorded the attendance of students at tutorials for a module. Suppose that we can, in some sense, regard the students as a sample from some population of students so that, for example, we can learn about the likely behaviour of next year's students by observing this year's. At the time I recorded the data we had had tutorials in Week 2 and Week 4. Let the probability that a student attends in both weeks be θ_{11}, the probability that a student attends in week 2 but not Week 4 be θ_{10} and so on. The data are as follows.

Attendance	Probability	Observed frequency
Week 2 and Week 4	θ_{11}	$n_{11} = 25$
Week 2 but not Week 4	θ_{10}	$n_{10} = 7$
Week 4 but not Week 2	θ_{01}	$n_{01} = 6$
Neither week	θ_{00}	$n_{00} = 13$

Suppose that the prior distribution for $(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ is a Dirichlet distribution with density proportional to

$$\theta_{11}^3 \theta_{10} \theta_{01} \theta_{00}^2.$$

(a) Find the prior means and prior variances of $\theta_{11},\ \theta_{10},\ \theta_{01},\ \theta_{00}$.

(b) Find the posterior distribution.

(c) Find the posterior means and posterior variances of $\theta_{11},\ \theta_{10},\ \theta_{01},\ \theta_{00}$.

(d) Using the R function `hpdbeta` which may be obtained from the Web page (or otherwise), find a 95% posterior hpd interval, based on the exact posterior distribution, for θ_{00}.

2. Suppose that we have J samples and, given the parameters, observation i in sample j is

$$y_{i,j} \sim N(\mu_j,\ \tau^{-1})$$

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$.
Let $\underline{\mu} = (\mu_1, \ldots, \mu_J)^T$, let $\underline{\bar{y}} = (\bar{y}_1, \ldots, \bar{y}_J)^T$, and let

$$S = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2,$$

where

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j}.$$

Show that $\underline{\bar{y}}$ and S are sufficient for $\underline{\mu}$ and τ.

3. We make n observations $y_1, \ldots, y_n$, which, given that values of parameters $\alpha,\ \beta$, are independent observations from a gamma(α, β) distribution. Show that the statistics $T_1,\ T_2$ are sufficient for $\alpha,\ \beta$ where

$$T_1 = \sum_{i=1}^{n} y_i \qquad \text{and} \qquad T_2 = \prod_{i=1}^{n} y_i.$$

4. Davies and Goldsmith (1972) give the following data on piston ring failures in steam-driven compressors. There were four identical compressors in the same compressor house, each oriented the same way, and each had three legs. The data give the number of failures in each leg of each compressor over a period of some years.

Compressor Number	North Leg	Centre Leg	South Leg
1	17	17	12
2	11	9	13
3	11	8	19
4	14	7	28

Let the number of failures in leg j (North: $j = 1$, Centre: $j = 2$, South: $j = 3$) of compressor i be $X_{i,j}$. Suppose that we regard the total number of failures, $N = 166$, as fixed and regard the numbers $X_{1,1}, \ldots, X_{4,3}$ as being an observation from a multinomial$(N, \theta_{1,1}, \ldots, \theta_{4,3})$ distribution. Suppose that our prior distribution for $\theta_{1,1}, \ldots, \theta_{4,3}$ is a Dirichlet$(a_{1,1}, \ldots, a_{4,3})$ distribution with $a_{i,j} = 2.0$ for all i, j.

(a) Find the posterior distribution for $\theta_{1,1}, \ldots, \theta_{4,3}$.

(b) Find the posterior mean for each of $\theta_{1,1}, \ldots, \theta_{4,3}$.

(c) For each of $\theta_{1,1}, \ldots, \theta_{4,3}$, find the symmetric 95% posterior interval, compare these intervals and comment.

Note: A symmetric 95% interval for θ is simply an interval (k_1, k_2) such that $\Pr(\theta < k_1) = \Pr(\theta > k_2) = 0.025$. You will need to use R to evaluate these intervals.

5. Ten measurements are made using a scientific instrument. Given the unknown value of a quantity θ, the natural logarithms of the measurements are independent and normally distributed with mean $\log \theta$ and known standard deviation 0.05.

Our prior distribution is such that $\log \theta$ has a normal distribution with mean 2.5 and standard deviation 0.5.

The logarithms of the measurements are as follows.

```
2.99   3.03   3.04   3.01   3.12   2.98   3.03   2.98   3.07   3.10
```

(a) Find the posterior distribution of $\log \theta$.

(b) Find a symmetric 95% posterior interval for $\log \theta$.

(c) Find a symmetric 95% posterior interval for θ.

(d) Find the posterior probability that $\theta < 20.0$.

6. Walser (1969) gave the following data on the month of giving birth for 700 women giving birth for the first time. The births took place at the University Hospital of Basel, Switzerland.

	Month	No. of births		Month	No. of births		Month	No. of births
1	January	66	5	May	64	9	September	54
2	February	63	6	June	74	10	October	51
3	March	64	7	July	70	11	November	45
4	April	48	8	August	59	12	December	42

We have unknown parameters $\theta_1, \ldots, \theta_{12}$ where, given the values of these parameters, the probability that one of these births takes place in month j is θ_j and January is month 1, February is month 2 and so on through to December which is month 12. Given the parameters, the birth dates are assumed to be independent.

Our prior distribution for $\theta_1, \ldots, \theta_{12}$ is a Dirichlet distribution with parameters $a_1 = a_2 = \cdots = a_{12} = 2$.

(a) Find the posterior distribution of $\theta_1, \ldots, \theta_{12}$.

(b) For each of $j = 1, \ldots, 12$, find the posterior mean of θ_j.

(c) For each of $j = 1, \ldots, 12$, find the posterior probability that $\theta_j > 1/12$ and comment on the results.

(d) Find the joint posterior distribution of $\theta_1, \theta_2, \tilde{\theta}_2$, where $\tilde{\theta}_2 = 1 - \theta_1 - \theta_2$.

Note: You may use R for the calculations but give the commands which you use with your solution.

7. Potatoes arrive at a crisp factory in large batches. Samples are taken from each batch for quality checking. Assume that each potato can be claasified as "good" or "bad" and that, given the value of a parameter θ, potatoes are independent and each has probability θ of being "bad."

(a) Suppose that m samples, each of fixed size n, are chosen and that the numbers of bad potatoes found are $x_1, \ldots, x_m$. Show that

$$s = \sum_{i=1}^{m} x_i$$

is sufficient for θ.

(b) Suppose that potatoes are examined one at a time until a fixed number r of bad potatoes is found. Let the number of potatoes examined when the r^{th} bad potato is found be y. This process is repeated m times and the values of y are $y_1, \ldots, y_m$. Show that

$$t = \sum_{i=1}^{m} y_i$$

is sufficient for θ.

(c) Suppose that we have a prior distribution for θ which is a beta(a, b) distribution. A two-stage inspection procedure is adopted. In Stage 1 potatoes are examined one at a time until a fixed number r of bad potatoes is found. The r^{th} bad potato found is the y^{th} to be examined. In Stage 2 a further n potatoes are examined and x of these are found to be bad.

i. Find the posterior distribution of θ after Stage 1.

ii. Find the posterior distribution of θ after Stage 1 and Stage 2.

Homework 4

Solutions to Questions **5, 6, 7** *of Problems 4 are to be submitted in the Homework Letterbox no later than 4.00pm on Monday April 20th.*

Solutions

1. (a) Prior distribution is Dirichlet(4,2,2,3).
So $A_0 = 4 + 2 + 2 + 3 = 11$.
The prior means are
$$\frac{a_{0,i}}{A_0}.$$
The prior variances are
$$\frac{a_{0,i}}{(A_0+1)A_0} - \frac{a_{0,i}^2}{A_0^2(A_0+1)}.$$
Prior means:
$$\begin{aligned}
\theta_{11}: \quad & \frac{4}{11} &=& \underline{0.3636} \\
\theta_{10}: \quad & \frac{2}{11} &=& \underline{0.1818} \\
\theta_{01}: \quad & \frac{2}{11} &=& \underline{0.1818} \\
\theta_{00}: \quad & \frac{3}{11} &=& \underline{0.2727}
\end{aligned}$$
Prior variances:
$$\begin{aligned}
\theta_{11}: \quad & \frac{4}{12\times 11} - \frac{4^2}{11^2\times 12} &=& \underline{0.019284} \\
\theta_{10}: \quad & \frac{2}{12\times 11} - \frac{2^2}{11^2\times 12} &=& \underline{0.012397} \\
\theta_{01}: \quad & \frac{2}{12\times 11} - \frac{2^2}{11^2\times 12} &=& \underline{0.012397} \\
\theta_{00}: \quad & \frac{3}{12\times 11} - \frac{3^2}{11^2\times 12} &=& \underline{0.016529}
\end{aligned}$$

(b) Posterior distribution is Dirichlet(4+25, 2+7, 2+6, 3+13). That is Dirichlet(29,9,8,16).

(c) Now $A_1 = 29 + 9 + 8 + 16 = 62$.
The posterior means are
$$\frac{a_{1,i}}{A_1}.$$
The posterior variances are
$$\frac{a_{1,i}}{(A_1+1)A_1} - \frac{a_{1,i}^2}{A_1^2(A_1+1)}.$$
Posterior means:
$$\begin{aligned}
\theta_{11}: \quad & \frac{29}{62} &=& \underline{0.4677} \\
\theta_{10}: \quad & \frac{9}{62} &=& \underline{0.1452} \\
\theta_{01}: \quad & \frac{8}{62} &=& \underline{0.1290} \\
\theta_{00}: \quad & \frac{16}{62} &=& \underline{0.2581}
\end{aligned}$$

Posterior variances:

$$\begin{aligned}
\theta_{11}: \quad & \frac{29}{63\times 62} - \frac{29^2}{62^2\times 63} &=& \underline{0.003952}\\
\theta_{10}: \quad & \frac{9}{63\times 62} - \frac{9^2}{62^2\times 63} &=& \underline{0.001970}\\
\theta_{01}: \quad & \frac{8}{63\times 62} - \frac{8^2}{62^2\times 63} &=& \underline{0.001784}\\
\theta_{00}: \quad & \frac{16}{63\times 62} - \frac{16^2}{62^2\times 63} &=& \underline{0.003039}
\end{aligned}$$

(d) Posterior distribution for θ_{00} is beta(16, $62-16$). That is beta(16,46).
Using the R command `hpdbeta(0.95,16,46)` gives $0.15325 < \theta_{00} < 0.36724$.

2. Let $N = \sum_{j=1}^{J} n_j$.
The likelihood is

$$\begin{aligned}
L &= \prod_{j=1}^{J}\prod_{i=1}^{n_j}(2\pi)^{-1/2}\tau^{1/2}\exp\left\{-\frac{\tau}{2}(y_{i,j}-\mu_j)^2\right\}\\
&= (2\pi)^{-N/2}\tau^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{i,j}-\mu_j)^2\right\}\\
&= (2\pi)^{-N/2}\tau^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{i,j}-\bar{y}_j+\bar{y}_j-\mu_j)^2\right\}\\
&= (2\pi)^{-N/2}\tau^{N/2}\exp\left\{-\frac{\tau}{2}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{i,j}-\bar{y}_j)^2+\sum_{j=1}^{J}n_j(\bar{y}_j-\mu_j)^2+2\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{i,j}-\bar{y}_j)(\bar{y}_j-\mu_j)\right]\right\}\\
&= (2\pi)^{-N/2}\tau^{N/2}\exp\left\{-\frac{\tau}{2}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{i,j}-\bar{y}_j)^2+\sum_{j=1}^{J}n_j(\bar{y}_j-\mu_j)^2\right]\right\}
\end{aligned}$$

since

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{i,j}-\bar{y}_j)(\bar{y}_j-\mu_j) = \sum_{j=1}^{J}\left\{(\bar{y}_j-\mu_j)\sum_{i=1}^{n_j}(y_{i,j}-\bar{y}_j)\right\} = 0.$$

Hence

$$L = (2\pi)^{-J/2}\tau^{J/2}\exp\left\{-\frac{\tau}{2}\left[S+\sum_{j=1}^{J}n_j(\bar{y}_j-\mu_j)^2\right]\right\}$$

in which the data only appear through S and $\underline{\bar{y}}$. Hence S and $\underline{\bar{y}}$ are sufficient for τ and $\underline{\mu}$.

3. Likelihood:

$$\begin{aligned}
L &= \prod_{i=1}^{n}\frac{\beta^\alpha y_i^{\alpha-1}e^{-\beta y_i}}{\Gamma(\alpha)}\\
&= \frac{\beta^{n\alpha}}{[\Gamma(\alpha)]^n}T_2^{\alpha-1}e^{-\beta T_1}\\
&= g(\alpha,\beta,T_1,T_2)h(\underline{y})
\end{aligned}$$

where $h(\underline{y}) = 1$.
So, by the factorisation theorem, T_1, T_2 are sufficient for α, β.

4.

5. (a) Prior mean: $M_0 = 2.5$
Prior precision:
$$P_0 = \frac{1}{0.5^2} = 4$$
Data precision:
$$n\tau = \frac{10}{0.05^2} = 4000$$
Posterior precision: $P_1 = 4 + 4000 = 4004$
Sample mean: $\bar{y} = 3.035$
Posterior mean:
$$M_1 = \frac{4 \times 2.5 + 4000 \times 3.035}{4004} = 3.0345$$
Posterior variance:
$$\frac{1}{4004} = 0.000250$$
Posterior distribution:
$$\underline{\log\theta \sim N(3.0345,\ 0.000250)}$$
(2 marks)

(b) Posterior interval for $\log\theta$: $M_1 \pm 1.96\sqrt{1/P_1}$
$$\underline{3.0035 < \log\theta < 3.0655}$$
(1 mark)

(c) Posterior interval for θ :
$$e^{3.0035} < \theta < e^{3.0655}$$
$$\underline{20.156 < \theta < 21.443}$$
(1 mark)

(d) Posterior probability:
$$\begin{aligned} \Pr(\theta < 20) &= \Pr(\log\theta < \log 20 = 2.9957) \\ &= \Phi\left(\frac{2.9957 - 3.0345}{\sqrt{0.000250}}\right) \\ &= \Phi(-2.45515) = \underline{0.0070} \end{aligned}$$
(1 mark)

6. (a) Prior density proportional to
$$\prod_{j=1}^{12} \theta_j^{2-1}$$
Likelihood proportional to

$$\prod_{j=1}^{12} \theta_j^{x_j}$$

Posterior density proportional to

$$\prod_{j=1}^{12} \theta_j^{x_j+2-1}$$

i.e. Dirichlet$(x_1+2,\ x_2+2,\ \ldots\ , x_{12}+2)$
Posterior distribution is Dirichlet(68, 65, 66, 50, 66, 76, 72, 61, 56, 53, 47, 44)

(2 marks)

(b) Posterior mean for θ_j is

$$\frac{x_j+2}{\sum(x_i+2)} = \frac{x_j+2}{\sum x_i + 24}$$

January	February	March	April	May	June
0.09392	0.08978	0.09116	0.06906	0.09116	0.10497
July	August	September	October	November	December
0.09945	0.08425	0.07735	0.07320	0.06492	0.06077

(1 mark)

(c) $\sum(x_j+2) = 724$
Marginal distribution for θ_j is beta$(x_j+2,\ 722-x_j)$

```
k<-1/12
prob<-1-pbeta(k,births+2,722-births)
```

Probability $\theta_j > 1/12$:

January	February	March	April	May	June
0.8357	0.7205	0.7630	0.0709	0.7630	0.9772
July	August	September	October	November	December
0.9322	0.5209	0.2650	0.1480	0.0286	0.0096

It seems very likely that some months have more than their "fair share" of births and some less. Of course there might have been something unusual about the period when the data were collected but, assuming there was not, then it seems very likely that the months are in fact different – even if we allowed for their different lengths. In particular, June and July seem to have high rates and April, November and December seem to have low rates.

Mark for reasonable comment.

(2 marks)

(d) If the posterior parameters are $a_{1,1}, a_{1,2}, \ldots, a_{1,12}$ then the joint posterior distribution of $\theta_1,\ \theta_2,\ \tilde{\theta}_2$ is Dirichlet$(a_{1,1}, a_{1,2}, A_1 - a_{1,1} - a_{1,2})$ where $A_1 = \sum a_{i,j}$.
Therefore the distribution is Dirichlet(68, 65, 591) .

(2 marks)

7. (a) Likelihood:

$$\begin{aligned} L &= \prod_{i=1}^{m} \binom{n}{x_i} \theta^{x_i}(1-\theta)^{n-x_i} \\ &= \left\{ \prod_{i=1}^{m} \binom{n}{x_i} \right\} \theta^{s}(1-\theta)^{nm-s} \\ &= g(\theta, s)h(\underline{x}) \end{aligned}$$

where $g(\theta, s) = \theta^s(1-\theta)^{nm-s}$ and $h(\underline{x}) = \prod_{i=1}^{m} \binom{n}{x_i}$.

Hence, by the factorisation theorem, s is sufficient for θ.

(2 marks)

(b) Likelihood:

$$\begin{aligned} L &= \prod_{i=1}^{m} \binom{y_i - 1}{r-1} \theta^{r}(1-\theta)^{y_i - r} \\ &= \left\{ \prod_{i=1}^{m} \binom{y_i - 1}{r-1} \right\} \theta^{mr}(1-\theta)^{t-mr} \\ &= g(\theta, t)h(\underline{y}) \end{aligned}$$

where $g(\theta, t) = \theta^{mr}(1-\theta)^{t-mr}$ and $h(\underline{y}) = \prod_{i=1}^{m} \binom{y_i - 1}{r-1}$.

Hence, by the factorisation theorem, t is sufficient for θ.

(2 marks)

(c) Prior density proportional to $\theta^{a-1}(1-\theta)^{b-1}$

i. Likelihood 1 proportional to $\theta^r(1-\theta)^{y-r}$
Hence posterior 1 proportional to $\theta^{a+r-1}(1-\theta)^{b+y-r-1}$
That is we have a $\underline{\text{beta}(a+r,\ b+y-r)}$ distribution.

(2 marks)

ii. Likelihood 2 proportional to $\theta^x(1-\theta)^{n-x}$
Hence posterior 2 proportional to $\theta^{a+r+x-1}(1-\theta)^{b+y+n-r-x-1}$
That is we have a $\underline{\text{beta}(a+r+x,\ b+y+n-r-x)}$ distribution.

(2 marks)