UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 4 June 2009 1:30 pm to 3:30 pm

PAPER 36

APPLIED BAYESIAN STATISTICS

*Attempt no more than **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Suppose we observe x positive responses out of m Bernoulli trials, each assumed conditionally independent given an unknown, common success chance θ. Our prior distribution for θ is Beta(a,b), with density $p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\ \theta^{a-1}\ (1-\theta)^{b-1};\ \ \theta \in (0,1)$.

(a) Derive the posterior distribution for θ given x.

(b) Show that the Jeffreys prior for θ is Beta(0.5,0.5). State the invariance property of Jeffreys prior distributions.

(c) We plan to observe a further n Bernoulli trials. If a and b above are positive integers, show that the predictive distribution for the future number of successes Y can be written as

$$p(y|n,x,m,a,b) = A \times B,$$

where

$$A = \frac{m+a+b-1}{m+n+a+b-1} \text{ and } B = \frac{\begin{pmatrix} y+x+a-1 \\ y \end{pmatrix}\begin{pmatrix} m+n-y-x+b-1 \\ n-y \end{pmatrix}}{\begin{pmatrix} m+n+a+b-2 \\ n \end{pmatrix}}.$$

(d) Another form of invariance property is as follows. Suppose we observe x out of m successes, and calculate the probability of observing y successes out of a further m trials. Compare this with the situation in which we had observed y out of m successes, and calculated the probability of x successes out of a further m trials. Then we should require that these two probabilities are the same. Show that this is true if $a = b = 1$.

(e) What other attractive predictive property does a Beta(1,1) prior have, as identified by Bayes?

(f) Can you interpret the form of B in $p(y|n,x,m,a,b)$ when $a = b = 1$?

2 Suppose you enter a town of unknown size whose trams are numbered consecutively from 1 to N. The first tram you see, assumed to be equally likely to be any of the trams, has number $y = 100$. We want to make inference on N.

(a) Show that the likelihood function for N is $\propto 1/N$; $N \geqslant y$.

(b) What is the maximum likelihood estimate of N?

(c) Suppose we assumed an improper discrete uniform prior distribution on the positive integers. What would be the posterior distribution for N given y? Why would this prior not be appropriate?

(d) Suppose we assumed a proper discrete uniform prior distribution on the integers 1 to M, where $M > y$. Find an exact expression for the posterior mean of N.

(e) Assume again a proper discrete uniform prior distribution on the integers 1 to M, where $M > y$. Making suitable approximations of sums by integrals, or otherwise, show that as M increases, $E[N|y] \times \log(M)/M \to 1$. Why would this lead us to expect our inferences on N to be very sensitive to the choice of M?

(f) Jeffreys suggested an improper prior $p(N) \propto 1/N$. Making suitable approximations of sums by integrals, or otherwise, show that $P(N \leqslant n|y) \approx 1 - y/n$, and hence that the posterior median is approximately $2y$.

(g) The following shows WinBUGS code for a version of Jeffreys's prior when the assumed maximum number of trams is 5000. Provide brief comments, focusing on the numbered lines of code, on what the code represents and why this will provide the desired analysis.

```
y <- 100
####################################
for(j in 1:5000){
reciprocal[j] <- 1/j
p.jeffreys[ j]<-reciprocal[j] / sum.recip  # (1)
}
sum.recip <- sum(reciprocal[])
N ~ dcat(p.jeffreys[])                     # (2)

y ~dcat(p[])
for(j in 1:5000){
  p[j] <-  step( N - j + 0.01) / N         # (3)
}
```

(h) When $y = 100$ we obtain the following results using Jeffreys's prior:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
N	408.7	600.4	4.99	102.0	197.0	2372.0	1001	10000

Why might the median be a better summary of the posterior than the mean?

3 Some classic mutagenicity assay data on salmonella features three plates that have been processed at each of six doses of quinoline, (recorded as μg per plate). The numbers of revertant colonies of TA98 Salmonella on each plate are shown below.

Dose level i	1	2	3	4	5	6
Dose x_i	0	10	33	100	333	1000
Plate 1	15	16	16	27	33	20
Plate 2	21	18	26	41	38	27
Plate 3	29	21	33	60	41	42

A certain dose-response curve is suggested by theory, so that for an observation Y_{ij} on the jth plate at the ith dose, we assume a Poisson model allowing for 'over-dispersion':

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\mu_{ij}) \text{ independently, given the } \mu_{ij}\text{'s} \\ \log \mu_{ij} &= \alpha + \beta \log(x_i + 10) + \gamma x_i + \lambda_{ij} \\ \lambda_{ij} &\sim \text{Normal}(0, \tau^2) \text{ independently.} \end{aligned}$$

(a) Just from examining the data by eye, why do you think an allowance for over-dispersion may be needed?

(b) In what way does this model allow for over-dispersion?

The model is fitted using the following WinBUGS code:

```
for(i in 1:doses) {
  for(j in 1:plates) {
  y[i,j] ~ dpois(mu[i,j])
  log(mu[i,j]) <- alpha + beta*log(x[i]+10) + gamma*x[i] + lambda[i,j]
  lambda[i,j] ~ dnorm(0.0, invtau2)
  }
}
alpha ~ dunif(-100,100)
beta  ~ dunif(-100,100)
gamma ~ dunif(-100,100)
tau~ dunif(0,100)
invtau2<-1/(tau*tau)
```

(c) How could the convergence be improved?

(d) Explain briefly the prior distributions given to the parameters, in particular why the standard Jeffreys prior is not given to variance parameter τ^2.

(e) How would you adapt the code if you wanted to fit a model with no overdispersion?

(f) The following table shows the DIC output based on 10000 iterations when fitting models with and without over-dispersion.

```
Dbar = post.mean of -2logL;
                                Dbar     pD      DIC
Model without over-dispersion  139.2    2.9    142.1
Model with over-dispersion     110.6   13.6    124.2
```

Interpret these results, in particular the `pD` column.

(g) How would you calculate standardised residuals around the fitted values for each *plate*? What would you be looking for and what would this procedure be checking?

(h) Suppose you wanted to check the underlying *dose-response* assumption by seeing if the predictions it would make matched the observed data. What replications might you make and how would you compare them with the observed data?

4

(a) Suppose we have two alternative models M_1 and M_2 with parameter vectors ψ_1 and ψ_2 respectively, and we are provided with prior distributions $p(\psi_i|M_i)$, sampling distributions $p_Y(y|\psi_i, M_i)$ for i=1,2, and a prior probability $p(M_1) = 1 - p(M_2)$. For an observation y, how would we find the posterior odds on model M_1?

In a (simplified version of a) micro-array experiment, we will make observations $Y_1, ..., Y_N$ which summarise the expression of N genes, where N is very large. Each Y_i is a standardised Normal variable with mean θ_i and variance 1, where θ_i is the true expression of gene i. If a gene i is 'negative', then $\theta_i = 0$. If gene i is 'positive', then θ_i is assumed to be a Normal variable with mean 0 and variance V, where V is assumed known. The proportion of 'positive' genes is denoted q, for the moment assumed known. We now observe a vector $y = (y_1, ..., y_N)$.

(b) For a positive gene, state the posterior mean of θ_i given y_i.

(c) Write down the predictive distribution for $Y_i|V$ for a positive gene i. Hence derive an expression for the posterior odds in favour of a gene being positive, as a function of y_i, V and q.

(d) Suppose now that q is unknown. Write down an expression for $p(y|q, V)$. Explain briefly how you might go about finding a maximum likelihood estimate for q?

(e) Suppose you have external information that q is around 10%, and is unlikely to be above 15%. In words, how might you tranform this information into a formal prior distribution?

(f) By introducing an indicator function or otherwise, provide rough WinBUGS code that will provide full posterior distributions for q and the θ_i's.

(g) Suppose you are not told the actual values comprising y, but only that 15% of the genes had an observed expression greater than 2. How might you estimate q?

END OF PAPER