UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 3 June, 2010 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

Derren claims he has extra-sensory perception (ESP) and can guess in advance how a fair coin will land (heads or tails) with a probability θ that is different from $\frac{1}{2}$. Let H_0 be the hypothesis that he does not have ESP, and H_1 be the hypothesis that he does have ESP.

(a) Define the prior odds on H_0.

(b) Suppose we have data y and a probability model that gives values for $p(y|H_0)$ and $p(y|H_1)$. Define the likelihood ratio and show how to obtain the posterior odds on H_0.

(c) Suppose the data comprises n flips of a coin, of which Derren got y correct. Assume that the prior distribution for θ is uniform over the interval $(0, 1)$. What is $p(y|H_1)$?

(d) By making a normal approximation to $p(y \mid H_0)$, show that the likelihood ratio (Bayes factor) is $\approx \sqrt{\frac{2n}{\pi}} \exp\left[-\frac{2}{n}\left(y - \frac{n}{2}\right)^2\right]$.

(e) Suppose that out of 10,000 flips, Derren gets 5150 right. In a classical statistical sense, is this statistically significant evidence against H_0?

(f) What, very approximately, is the Bayes factor between H_0 and H_1? How do you explain any difference between this and the 'classical' result?

(g) Informally, what might be a more reasonable prior for θ under H_1?

(h) Even if further evidence gives a Bayes factor in favour of H_1, do you think the posterior odds on H_0 should necessarily be less than 1?

[*A Beta*(a, b) *distribution has density* $p(\theta|\, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \theta^{a-1}\,(1-\theta)^{b-1}\,; \quad \theta \in (0, 1)$.]

2

Let Y_i be the number of cases of leukaemia over a 10-year period in area i, $i = 1, \ldots, I$. Assume that $Y_i \sim \text{Poisson}(\lambda_i E_i)$, where E_i is the expected number of cases based on the age and sex of the population of area i and assuming a constant rate over the whole country. λ_i is known as the Standardised Incidence Rate (SIR), which expresses the leukaemia risk.

(a) Show that the Jeffreys prior is $p_J(\lambda_i) \propto 1/\sqrt{\lambda_i}$.

(b) After observing a count y_i, what is the posterior distribution for λ_i assuming the Jeffreys prior?

(c) Show how the Jeffreys prior can be approximately expressed as a proper Gamma distribution.

(d) Someone wants to create a 'league table' of areas ranked according to their leukaemia risk, and to identify the area with the highest risk. Briefly describe how you would write a simulation program that would give posterior distributions for the ranks of the areas, and the probability that each area had the highest risk. (You may express this using rough BUGS code if you wish, but the syntax does not need to be correct.)

(e) In fact a certain amount of variability, above that due to age and sex, between areas is inevitable, and so it is suggested that a prior distribution for each λ_i with mean 1 and standard deviation 0.25 might be reasonable. Show how to express this as a Gamma distribution.

(f) Using this new prior distribution, what is the posterior distribution for each λ_i?

(g) What effect would you expect this new prior distribution to have on the distributions of the ranks?

(h) Someone suggests using the $p(\lambda_i > 1 \,|\, y_i)$ as a measure of whether area i has a seriously increased risk of leukaemia. Why do you think they suggest this, and can you see any problems with it when an informative prior distribution is used? What might be a better measure?

[*A Gamma*(a, b) *distribution has density* $p(\lambda|\, a, b) = \frac{b^a}{\Gamma(a)}\, \lambda^{a-1}\, e^{-\lambda b}$; $\lambda \in (0, \infty)$, *with mean* a/b *and variance* a/b^2 .]

3

Let $D(\theta) = -2 \log p(y \mid \theta)$ be the deviance arising from data $y = y_1, \ldots, y_n$, with each y_i assumed independently drawn from a specified distribution with a p-dimensional parameter vector θ.

(a) Show that the maximum likelihood estimate $\hat{\theta}$ minimises the deviance $D(\theta)$.

(b) Assume that θ has a locally uniform (non-informative) prior distribution $p(\theta) \propto$ constant. Assuming the standard asymptotic properties of $\hat{\theta}$, show (non-rigorously) that θ has posterior distribution

$$\theta \sim \text{Normal}_p \left(\hat{\theta}, \left[\frac{1}{2} \left. \frac{\partial^2 D}{\partial \theta^2} \right|_{\hat{\theta}} \right]^{-1} \right) .$$

(c) By expanding $D(\theta)$ about $\hat{\theta}$, show that $D(\theta) \approx D(\hat{\theta}) + X_p^2$, where X_p^2 indicates a variable with a χ_p^2 distribution.

(d) Define p_D, the effective number of parameters. Show that $p_D \approx p$ when assuming a locally uniform prior for θ.

(e) Define the Deviance Information Criterion (DIC) and the Akaike Information Criterion (AIC), and show that they will be approximately equivalent with a locally uniform prior for θ.

(f) When using MCMC and a locally uniform prior for θ, suggest two ways of estimating the minimum deviance $D(\hat{\theta})$.

(g) It has been suggested that half the posterior variance of the deviance should be used as the effective number of parameters. Why might this be appropriate?

[*A* χ_p^2 *distribution has mean* p *and variance* $2p$.]

4

A sample of 106 children in Gambia were immunised against Hepatitis B at a baseline visit and then followed up at 3 additional clinic visits. Their level of immunity is measured by their 'titre'. Let Y_{ij} be the log(titre) of child i at clinic visit j at time t_{ij}, where t_{ij} is the time since immunisation. Y_{ij} is assumed to be drawn from a Normal(μ_{ij}, σ^2) distribution. We assume that each child's expected log(titre) μ_{ij} changes linearly with $\log(t_{ij})$, and also depends on the child's baseline log-titre y_{0i}, and has a different intercept for each child, so that $\mu_{ij} = \alpha_i + \beta \log(t_{ij}) + \gamma y_{0i}$, and $\alpha_i \sim$ Normal(δ, τ^2).

This is Model 1 and is fitted using the following WinBUGS code:

```
for    ( i in 1:106 ) {
   for ( j in 1:3   ) {
       y  [i, j]    ~ dnorm ( mu [i, j], invsigma2 )
       mu [i, j] ) <- alpha [ i ] + beta*log ( time [i, j]) + gamma*y0 [ i ]
  }
  alpha [ i ] ~ dnorm ( delta, invtau2 )
}

invsigma2 ~ dgamma (0.001, 0.001 )
beta      ~ dunif  (-100,  100 )
gamma     ~ dunif  (-100,  100 )
delta     ~ dunif  (-100,  100 )
tau       ~ dunif  (   0,  100 )
invtau2  <- 1 / ( tau*tau )
```

(a) Explain briefly what will be effect of assuming the α_i's are drawn from a common prior distribution.

(b) How might the convergence be improved?

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

(d) Why might it be reasonable to assume the baseline log-titre y_{0i} is an observation from a distribution that is Normal(μ_{0i}, σ^2), where μ_{0i} is the true baseline titre?

(e) Consider Model 2 in which the regression model is changed to

$$\mu_{ij} = \alpha_i + \beta \log(t_{ij}) + \gamma \mu_{0i},$$

and $\mu_{0i} \sim$ Normal(θ, ψ^2). Why would you want to consider such a model? Draw a rough directed graph for the whole of Model 2. How would you adapt the code if you wanted to fit Model 2? [Do not worry about correct syntax.]

(f) Model 2 gave the following output

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta	-1.064	0.1352	0.001286	-1.329	-1.065	-0.7955	1001	10000
gamma	1.023	0.1145	0.0106	0.7915	1.014	1.231	1001	10000

In Model 3, we fix $\beta = -1$, $\gamma = 1$. Why might this be a reasonable assumption?

(g) The following table shows the DIC output based on 10000 iterations when fitting the models 2 and 3.

```
Dbar = post.mean of -2logL;

            Dbar    pD     DIC
Model 2     1128.1  143.6  1271.7
Model 3     1128.3  141.5  1269.8
```

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

(h) Explain why Model 3 could be interpreted as implying that the fraction of titre after time t decreases as $1/t$.

END OF PAPER