UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 9 June, 2011 9:00 am to 11:00 am

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

John has a coin with unknown probability θ of coming up heads. He tells me he has flipped it until it first came up heads, which happened on flip $y+1$, after y tails had occurred.

(a) Why is y an observation from a geometric distribution $p(y|\theta) = (1-\theta)^y\theta$, $y = 0, 1, 2, ..$?

(b) Define the Jeffreys prior $p_J(\theta)$ for a scalar parameter θ and derive its form for this geometric distribution. Can you express this distribution as a limiting member of a parametric family?

(c) Assuming the Jeffreys prior $p_J(\theta)$, what is the posterior distribution for θ?

(d) Prove in general that, for any 1-1 function $\mu(\theta)$, $p_J(\mu) = p_J(\theta)\left|\frac{d\theta}{d\mu}\right|$.

(e) The mean of the geometric distribution is $\mathrm{E}[Y] = \mu = (1/\theta) - 1$. Express the sampling density $p(y|\mu)$ in terms of μ, and find the form for $p_J(\mu)$.

(f) Suppose John now tells me that he lied about flipping the coin until he got a head: he just flipped the coin $y+1$ times, got y tails and 1 head in some order and then got bored. We can assume that y is an observation from a Binomial distribution with index $y+1$ and parameter $1-\theta$, so that $p(y|\theta) = \begin{pmatrix} y+1 \\ y \end{pmatrix}(1-\theta)^y\theta$. The Jeffreys prior for the Binomial model is known to be a Beta$(\frac{1}{2}, \frac{1}{2})$ distribution. What is the posterior distribution for θ under this Binomial model and Jeffreys prior?

(g) Compare the posterior means of θ under the geometric model with its Jeffreys prior and the Binomial model with its Jeffreys prior.

(h) State the likelihood principle.

(i) Do the above conclusions obey the likelihood principle?

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

2

Suppose we have an observation y assumed to arise from a density $p(y|\theta)$, which can be transformed to a member of the natural exponential family $p(t|\phi) = e^{\alpha(t)} e^{\phi t - k(\phi)}$, where $t = t(y)$ and $\phi = u(\theta)$ are the natural observation and parameter.

(a) Describe a family of prior densities that is conjugate to $p(t|\phi)$, and show why the family is conjugate.

(b) Assume that y is a single Bernoulli observation with $p(y|\theta) = \theta^y (1-\theta)^{1-y}$. Show that this can be transformed to the natural exponential family and provide expressions for t, ϕ and $k(\phi)$.

(c) Find an expression proportional to the conjugate prior density for a Bernoulli observation expressed in terms of ϕ.

(d) Find the corresponding posterior distribution for ϕ after observing a Binomial observation $r \sim \text{Bin}(n, \theta)$ with $p(r|n, \theta) = \begin{pmatrix} n \\ r \end{pmatrix} \theta^r (1-\theta)^{n-r}$.

(e) Show that this posterior for ϕ is equivalent to a Beta posterior for θ.

(f) Suppose we observe another Binomial observation $\tilde{r} \sim \text{Bin}(m, \theta)$, assumed conditionally independent of r given θ. What form does the posterior distribution $p(\theta|\tilde{r}, r)$ take?

(g) Suppose, in a general parametric model, we wish to derive the predictive distribution for a new observation $\tilde{y}$ assumed conditionally independent of y given θ. By expanding the joint density of $p(\tilde{y}, \theta|y)$ in two different ways, or otherwise, prove that, for any θ,

$$p(\tilde{y}|y) = \frac{p(\tilde{y}|\theta)\ p(\theta|y)}{p(\theta|\tilde{y}, y)}.$$

(h) Use the above expression to derive the predictive distribution for a subsequent Binomial observation $\tilde{r}$.

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

3

An insurance company takes a sample of $n = 35$ customers and adds up the number of claims each has made in the past year to give $y_1, .., y_n$. These form the following distribution, where $m_k = \sum_i I(y_i = k)$ is the number of $y's$ that equal k

k	0	1	2	3	4	5
m_k	10	5	10	7	2	1

The total number of claims in the sample is $\sum_k km_k = 59$ and the average number of claims per customer is $\sum_k m_k/35 = 1.69$. The company wants to use this sample to estimate the mean number of claims μ per customer in the last year. They know that the mean number of claims per customer varies from year to year, and in past years this mean has varied around 2 claims per year with an approximate standard deviation of 0.5.

(a) If we assume a Poisson sampling model $p(y_i|\mu) = e^{-\mu}\mu^{y_i}/y!$ with mean μ, why might it be convenient to assume a Gamma prior distribution for μ?

(b) What specific Gamma distribution would be appropriate?

(c) What is the posterior distribution for μ assuming this sampling model and prior?

(d) Someone now points out that there seem to be rather a lot of customers making no claims, which may make the Poisson assumption inappropriate. Let $G = m_0 m_2/m_1^2$. For this data-set, calculate G. In data that genuinely come from a Poisson distribution, what (approximately) would it be reasonable to expect as a value for G? Why might G be a good statistic to use as a checking function?

(e) How might we use replicate data to check whether our observed value of G is unusual under a Poisson assumption? [You can use rough BUGS code - it does not have to be syntactically correct]

(f) On the basis of this analysis we decide the Poisson model is not realistic, and use the following BUGS code for what is known as the 'zero-inflated Poisson' model.

```
for ( i in 1:35){
   y[i]      ~ dpois(mean[i])     (1)
   mean[i] <- (1-group[i])*mu    (2)
   group[i] ~ dbern(p)           (3)
   }
```

Describe the model expressed by the code in lines (1) to (3), and interpret the parameters p (`p`) and μ (`mu`).

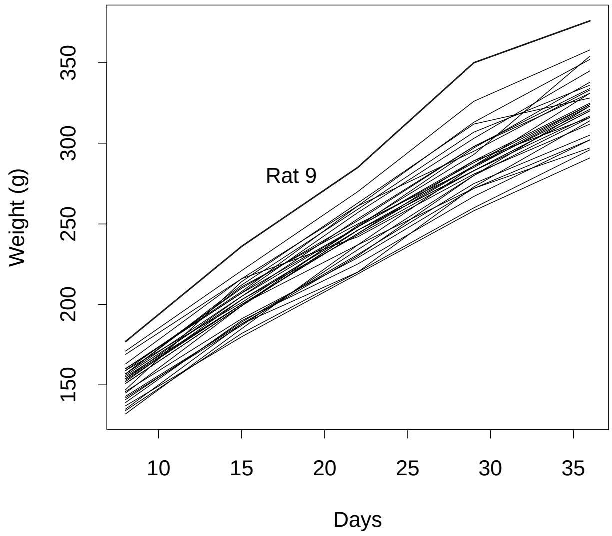
(g) In terms of p and μ, what is the mean of the distribution being fitted? In terms of p and μ, what is the overall expected proportion of zero-claims?

(h) Briefly, how might you use historical data to assess a joint prior distributions for p, μ?

[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.]

4

The figure below shows measurements of the weights in grams of 30 rats taken at 8, 15, 22, 29 and 36 days of age.



Let Y_{ij} be the weight of rat i at weighing j at age x_j days. The Y_{ij} are assumed to be drawn independently from a Normal(μ_{ij}, σ^2) distribution, and each rat's expected weight μ_{ij} changes linearly with age x_j and has a different intercept and gradient for each rat, so that $\mu_{ij} = \beta_{i1} + \beta_{i2}(x_j - \overline{x})$, where $\overline{x} = 22$ is the mean age at weighing.

We first assume a hierarchical model (Model 1) in which the intercepts have a distribution $\beta_{i1} \sim$ Normal$(\mu_{\beta_1}, \tau_1^2)$, and the gradients are dependent on the intercepts, with distribution $\beta_{i2}|\beta_{i1} \sim$ Normal$(\mu_{\beta_2} + \gamma(\beta_{i1} - \mu_{\beta_1}), \tau_2^2)$. Given the hyper-parameters $(\mu_{\beta_1}, \mu_{\beta_2}, \gamma, \tau_1, \tau_2)$, (β_{i1}, β_{i2}) are independent across rats. The full model is described using the following WinBUGS code:

```
for( i in 1 : 30 ) {
  for( j in 1 : 5 ) {
    y[i , j]    ~ dnorm(mu[i , j], invsigma2)
    mu[i , j]   <- beta[i , 1] + beta[i , 2] * (x[j]-mean(x[]))
  }
  beta[i , 1]    ~ dnorm(mu.beta[1], invtau2[1])
  mean.beta[i,2]<- mu.beta[2]+gamma*(beta[i,1] - mu.beta[1])
  beta[i , 2]    ~ dnorm(mean.beta[i,2], invtau2[2])
}
for(i in 1:2){
  mu.beta[i] ~ dunif(-1000,1000)
  invtau2[i] <- 1 / (tau[i]*tau[i])
  tau[i]     ~ dunif(0,100)
}
gamma     ~ dunif(-100,100)
invsigma2 ~ dgamma(0.001, 0.001)
sigma     <- 1 / sqrt(invsigma2)
```

(a) How would you interpret the intercept β_{i1}? Examining the data by eye, would you say it might be reasonable to model each gradient as dependent on the corresponding intercept? What is the interpretation of γ?

(b) What advantage should there be in centring the covariate around its mean? Briefly, what advantage might there be in fitting this conditional model rather than directly modelling the intercept and gradient as coming from a bivariate normal distribution?

(c) Explain briefly the prior distributions given to the parameters, in particular why the prior given to the random-effects variance parameters τ_1^2 and τ_2^2 is different from that given to σ^2.

(d) Analysing the data assuming Model 1 gave the following output:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
mu.beta[1]	242.7	2.7	0.04	237.2	242.7	248.0	1001	10000
mu.beta[2]	6.19	0.11	0.002	5.97	6.19	6.41	1001	10000
gamma	0.024	0.007	0.0001	0.011	0.024	0.037	1001	10000
sigma	6.09	0.47	0.008	5.26	6.06	7.11	1001	10000
tau[1]	14.87	2.14	0.023	11.35	14.66	19.71	1001	10000
tau[2]	0.406	0.091	0.002	0.241	0.400	0.599	1001	10000

How would you interpret the results for γ? Is the Monte Carlo error sufficiently small to accept these Monte Carlo estimates of the posterior means?

(e) Someone looks at the plot and suggests that maybe it is reasonable to assume all the gradients are the same. What parameter constraints would this be equivalent to? Would the output from Model 1 suggest this was reasonable?

(f) You decide to use a prior model assuming a constant gradient (model 2). Write rough BUGS code to do this.

(g) The following table shows the DIC output based on 10000 iterations when fitting the models 1 and 2.

	Dbar	pD	DIC
Model 1 (random gradients)	966.8	51.5	1018.350
model 2 (equal gradients)	1058.3	30.4	1088.720

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

(h) Rat 9 could be an outlier. Briefly, how might you check whether this is the case?

(i) How might you adapt the code for Model 1 to allow some rats to have outlying gradients?

(j) Suppose you take a single weighing of a new rat and it weighs 180 gms at 8 days, a little heavier than rat 9 at that age. How would you predict what its growth will be? Would you expect the predicted growth to stay above that of rat 9?

END OF PAPER