Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2010–2011**

MAS6004 Inference **3 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
Marks will be awarded for your best ***five*** *answers. Total marks 100.*

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 In an investigation into the accuracy of a new measurement instrument, n measurements are taken of the mass of a standard sample of mass 1000 grammes. The values obtained (in grammes), $x_1, \ldots, x_n$, can be modelled as exchangeable normal random variables with known mean $\mu = 1000$ and unknown variance σ^2.

(i) Show that the Inverse Gamma family of distributions is a conjugate family of priors for σ^2 in this situation, and that the parameters are updated as follows:

$$\begin{aligned} d &\rightarrow d + n/2, \\ a &\rightarrow a + \sum (x_i - \mu)^2/2. \end{aligned}$$

(6 marks)

Recall that σ^2 has an inverse gamma distribution with parameters d and a, written $IG(d, a)$, if it has density

$$f(\sigma^2) = \frac{a^d (\sigma^2)^{-(d+1)}}{\Gamma(d)} \exp\left(-\frac{a}{\sigma^2}\right),$$

for $\sigma^2 > 0$, and that provided $d > 2$, then

$$E(\sigma^2) = \frac{a}{d-1}$$

and

$$Var(\sigma^2) = \frac{a^2}{(d-1)^2(d-2)}.$$

(ii) Before taking any measurements, a scientist has priors beliefs summarised by $E(\sigma^2) = 0.01, Var(\sigma^2) = (0.005)^2$. Obtain a suitable conjugate prior representing these beliefs. ***(4 marks)***

(iii) An initial measurement is taken with the new instrument, and gives $x_1 = 1000.01$. What is the posterior distribution for σ^2, for the scientist in part (ii)? ***(2 marks)***

(iv) A further 9 measurements are taken, and the whole series, including the measurement from part (iii), can be summarised by $\sum_{i=1}^{10} (x_i - 1000)^2 = 0.025$. Given these measurements, calculate the posterior mean and variance for σ^2, for the scientist in part (ii). ***(4 marks)***

(v) The new instrument is one of a collection, all manufactured in a similar way. The properties of these instruments are considered to be exchangeable, and similar kinds of test data are available for each of them. Explain briefly how a hierarchical model could be used to describe the relationships between the properties of the instruments and the data. ***(4 marks)***

2 (i) Define $X \sim Binomial(n, \theta)$ and $Y \sim Binomial(m, \theta)$ to be conditionally independent, conditional on the value of θ, and let θ have a $Beta(a, b)$ prior distribution. Write down (a) the posterior distribution for θ given X, and (b) the predictive distribution for Y given X. (You do not need to *derive* these results.) ***(2 marks)***

(ii) Two gamblers are interested in the long-run probability of getting heads, θ, when a particular coin is tossed repeatedly. They agree that their beliefs are symmetric around $\theta = 1/2$. Gambler 1 has prior variance is $1/20$ for θ, and Gambler 2 has prior variance $1/100$. In each case, obtain suitable Beta distributions to represent these prior beliefs. ***(4 marks)***

(iii) A series of 8 tosses of the coin is observed (by both gamblers), and produces 4 heads and 4 tails. For each gambler, obtain the posterior distribution for θ and the predictive probabilities that (a) the next toss of the coin gives a head, and (b) the next 4 tosses of the coin all give heads. Comment briefly on how the gamblers' predictions are influenced by their priors.
(11 marks)

(iv) Without further calculation, explain what would happen to these predictive probabilities after the gamblers had seen a large number of tosses of the coin, of which a proportion h were heads. ***(3 marks)***

3 An investigation into the numbers of cases of a certain rare disease in 2009, in various towns and cities in the UK, gave the following figures.

City/town	Population (thousands)	Number of cases
Wolverhampton	251	20
Derby	229	13
Norwich	174	11
Oxford	143	7
St Helens	103	6
Crawley	101	5

The Winbugs code below defines a possible model for the occurrence of the disease.

```
model {
for (j in 1:6){
lambda[j] <- alpha[j] *theta[j]
x[j] ~ dpois(lambda[j])
theta[j] ~ dgamma(psi,rho) }
psi ~ dgamma(0.001,0.001)
rho ~ dgamma(0.001,0.001)
}
list(alpha=c(251,229,174,143,103,101),x=c(20,13,11,7,6,5))
```

(i) Draw a Directed Acyclic Graph (DAG) to illustrate the model represented by the above code. ***(5 marks)***

(ii) Explain the structure and assumptions of the above model, and the meaning of the variables, in a form suitable for a Bayesian statistician who is not familiar with Winbugs. ***(6 marks)***

(iii) The quantities

```
mu <- psi/rho
cv <- 1/sqrt(psi)
```

(in Winbugs notation) represent the mean and the coefficient of variation of a particular distribution. Explain their interpretations. ***(4 marks)***

(iv) For the city of Stoke-on-Trent, population 259,000, the number of cases in 2009 is unknown. Write down the additional Winbugs code necessary to sample from the posterior distribution for this quantity, explaining any additional assumptions you are making. ***(5 marks)***

4 An economic model has been constructed to predict the cost per patient of a new treatment for osteoporosis. In addition to the price of the drug, the model includes costs of hospital visits, nursing care, and any necessary surgery. The model has two uncertain inputs:
x: the time in days until the patient's first hip fracture
y: the number of days of nursing home care required, if the patient suffers a fracture.

M is defined to be the expected cost per patient, and P_1 is the probability that a patient's cost will exceed 100,000 pounds.

The following distributions are assumed for x and y:

$$\begin{aligned} x &\sim exponential(rate = 1/30), \\ y &\sim N(20, 25). \end{aligned}$$

The model is implemented in R using a user-defined function `cost(x,y)`. If `x` and `y` are vectors, then `cost(x,y)` will return the appropriate vector output. Some output from the R session is given below.

```
> n<-1000 > x<-rexp(n,1/30)
> y<-rnorm(n,20,5)
> c1<-cost(x,y)
> mean(c1)
[1] 55538.2
> var(c1)
[1] 364736611
> sum(c1>100000)
[1] 72
```

(i) Estimate M and P_1, giving 95% confidence intervals for both. What would you expect to happen to your 95% confidence interval for P_1 if `n` were changed to 100000? ***(6 marks)***

 Question 4 continued on next page

4 (continued)

(ii) A modification to the listing is made:

```
> u.1<-runif(500,0,1)
> u.2<-1-u.1
> x<-c( -30*log(1-u1) , -30*log(1-u2) )
> y<-rnorm(n,20,5)
> c1<-cost(x,y)
> mean(c1)
[1] 60483.1
> var(c1)
[1]  341901123
> cor(c1[1:500],c1[501:1000])
[1] -0.511
```

(a) Regarding the first three lines, name the two techniques that have been used to produce `x`. Give the motivation for this modification to the code. ***(3 marks)***

(b) Based on the new output, calculate a 95% confidence interval for M. ***(6 marks)***

(iii) An alternative distribution is proposed for y: the $Gamma(10, 0.5)$ distribution, with density function

$$f(y) = \frac{0.5^{10}}{\Gamma(10)} y^9 \exp(-0.5y),$$

for $y > 0$.

(a) If the original R analysis generated output values output $c_1, \ldots, c_{1000}$ from input values $x_1, \ldots, x_{1000}$ and $y_1, \ldots, y_{1000}$, give a formula for the Monte Carlo estimate of M, corresponding to the new distribution of y, in terms of $c_1, \ldots, c_{1000}$ and $y_1, \ldots, y_{1000}$, which could be calculated without doing any further evaluations of the function `cost`. ***(3 marks)***

(b) Explain how you would calculate a 95% confidence interval for M in this case.

(2 marks)

5 Given observations (x_i, y_i) for $i = 1, \ldots, 100$, the following model is considered:

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, 100$. The data are plotted below.

=4in q2fig$_2$011.*eps*

In an R session, with $(y_1, \ldots, y_{100})$ and $(x_1, \ldots, x_{100})$ stored under the variable names `y` and `x` respectively, the least squares estimate of β is obtained with the commands

```
lm1<-lm(y~x)
lm1$coefficients[2]
```

The least squares estimator $\hat{\beta}$ of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{100}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{100}(x_i - \bar{x})^2}.$$

5 (continued)

(i) Below is some output from an R session.

```
k<-rep(0,1000)
for(i in 1:1000){
z<-sample(x,size=100,replace=F)
lm2<-lm(y~z)
k[i]<-lm2$coefficients[2] }

lm1<-lm(y~x)
a<-lm1$coefficients[2]
sum(abs(k)<abs(a))
[1] 998
```

(a) State the null hypothesis being tested, explain the procedure that has been used to conduct the test, and give the result of the test. Why might this procedure be preferable to a test derived from the assumption that $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, 100$? ***(6 marks)***

(b) Derive a simpler function of the data that could be used in definitions of `k[i]` and `a`, without changing the result (assuming the same seed was used in each case). ***(4 marks)***

(c) If there were only 5 observations rather than 100, what is the smallest possible p-value that could be obtained using this procedure? ***(3 marks)***

(ii) Some further R code from the same session is given below.

```
k<-rep(0,1000)
for(i in 1:1000){
a<-sample(c(1:100),size=100,replace=T)
r<-y[a]
z<-x[a]
lm2<-lm(r~z)
k[i]<-lm2$coefficients[2] }

quantile(k,c(0.05,0.95))
        5%        95%
0.06825594 0.31096000
```

Give the name and a brief description of the statistical procedure that has been used here. What do the two numbers in the last line represent? ***(3 marks)***

5 (continued)

(iii) Given observations (x_i, y_i) for $i = 1, \ldots, 100$, consider the model

$$y_i = \beta x_i + \varepsilon_i,$$

with $\varepsilon_i \sim t_1$, for $i = 1, \ldots, 100$, so that each error has a Student-t distribution with 1 degree of freedom. Explain carefully the steps required to perform a Monte Carlo test of size 0.01 of the hypothesis $\beta = 0$, using the least squares estimator of β as a test statistic. You are not required to provide any R code. ***(4 marks)***

6 A sample of independent random observation $X = \{X_1, \ldots, X_n\}$ are drawn from the following mixture distribution:

$$X_i \sim \begin{cases} N(\mu, \sigma^2) & \text{with probability } \phi \\ Exponential(\text{rate} = \lambda) & \text{with probability } 1 - \phi \end{cases}$$

Define $\theta = (\mu, \sigma^2, \lambda, \phi)$ to be the vector of the four unknown parameters. The corresponding 'missing' variables $Y = \{Y_1, \ldots, Y_n\}$ are defined as follows:

$$Y_i = \begin{cases} 1 & \text{if } X_i \text{ is drawn from } N(\mu, \sigma^2), \\ 0 & \text{if } X_i \text{ is drawn from } Exponential(\text{rate}= \lambda). \end{cases}$$

However, X_1 and X_2 are observed to be negative, taking distinct values, and so $Y_1 = 1$ and $Y_2 = 1$ are known.
Sufficient statistics for $\theta = (\mu, \sigma^2, \lambda, \phi)$ are

$$\begin{aligned} S_1(X, Y) &= \sum_{i=1}^{n} Y_i, \\ S_2(X, Y) &= \sum_{i=1}^{n} Y_i X_i, \\ S_3(X, Y) &= \sum_{i=1}^{n} Y_i X_i^2, \\ S_4(X, Y) &= \sum_{i=1}^{n} X_i \end{aligned}$$

(i) Write down the complete data log-likelihood function for $\theta = (\mu, \sigma^2, \lambda, \phi)$ given both X and Y. ***(3 marks)***

(ii) Derive expressions for $E\{S_1(X, Y)|\theta\}$, $E\{S_2(X, Y)|\theta\}$, $E\{S_3(X, Y)|\theta\}$ and $E\{S_4(X, Y)|\theta\}$. ***(5 marks)***

(iii) Let $\theta_{old} = (\mu_{old}, \lambda_{old}, \sigma^2_{old}, \phi_{old})$ denote an initial estimate θ.

(a) Derive an expression for $p_i = E(Y_i|X, \theta_{old}, Y_1, Y_2)$, for $i > 2$. ***(4 marks)***

(b) Derive expressions for $E\{S_1(X, Y)|X, Y_1, Y_2, \theta_{old}\}$, $E\{S_2(X, Y)|X, Y_1, Y_2, \theta_{old}\}$, $E\{S_3(X, Y)|X, Y_1, Y_2, \theta_{old}\}$ and $E\{S_4(X, Y)|X, Y_1, Y_2, \theta_{old}\}$. You may leave your expressions in terms of p_i. ***(5 marks)***

(iv) Using your results from parts (ii) and (iii), apply one iteration of the EM algorithm to obtain expressions for an improved estimate of θ. ***(3 marks)***

End of Question Paper