Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2014–2015**

MAS472 Computational Inference **2 hours**

Candidates may bring to the examination a calculator that conforms to University regulations.
*Marks will be awarded for your best **three** answers. Total marks 90.*

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 (i) Let X be a random variable with the triangular density function

$$f_X(x) = \begin{cases} x & \text{for } 0 \leq x \leq 1, \\ 2 - x & \text{for } 1 < x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the cumulative distribution function of X. ***(4 marks)***

(b) Using your result from part a), explain how to generate a random value of X given a uniform random number using the inversion method. ***(6 marks)***

(ii) Suppose it is desired to generate a random variable X where X has the following density function:

$$f_X(x) = \begin{cases} \frac{3}{2}(1 - x^2) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Explain carefully how to generate a random value of X using rejection sampling, with a uniform envelope density function. If the first candidate value is $Y = 0.5$, given the value $U = 0.8$ from the $U[0, 1]$ distribution, determine whether 0.5 is accepted as a random sample from the distribution of X. ***(12 marks)***

(b) Using the same envelope density function, what is the expected number of candidate random variables Y required to obtain a single random X? ***(2 marks)***

(c) Consider the alternative envelope density function

$$g(y) = \begin{cases} 2(1 - y) & \text{for } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Derive the expected number of candidate random variables Y required to obtain a single random X with this new envelope function. ***(6 marks)***

[Hint: $(1 - x^2) = (1 - x)(1 + x)$**.]**

2 (i) A mathematical model has been developed to predict the total profit (in £,000,000) made by development of an oil field in the North Sea. The company considering development of the field wishes to consider uncertainty in the model output (the total profit) that results from uncertainty in model inputs. The uncertain inputs are θ the total volume (million barrels) of all the hydrocarbons in the oil field, and ϕ the proportion of the hydrocarbons which are crude oil. Uncertainty about these two quantities is described by the following distributions:

$$\begin{aligned} \log\theta &\sim N(3, 5^2), \\ \phi &\sim Beta(4, 2), \end{aligned}$$

with θ and ϕ independent. The model is represented by the deterministic function $c(\theta, \phi)$, and the expected profit is defined as

$$M = \int_0^1 \int_0^\infty c(\theta, \phi) p(\theta) p(\phi) d\theta d\phi,$$

with $p(\theta)$ and $p(\phi)$ the density functions of θ and ϕ respectively.

It is not possible to calculate M explicitly, and so M is estimated using the following R code. The function $c(\theta, \phi)$ is implemented in R using a user-defined function `profit(theta,phi)`. If `theta` and `phi` are vectors, then `profit(theta,phi)` will return the appropriate vector output. Some output from the R session is given below.

```
> theta <- exp(rnorm(400,3,5))
> phi <- rbeta(400,4,2)
> a <- profit(theta,phi)
> mean(a)
[1] 0.243
> var(a)
[1] 8.176
```

(a) Estimate the expected profit M, and give a 95% confidence interval for M. ***(4 marks)***

[Note: `qnorm(0.975) = 1.96`**]**

(b) If the company wants to obtain another 95% confidence interval for M that is no wider than 0.1, how many sampled pairs (θ, ϕ) would they need? ***(4 marks)***

2 (continued)

(c) An alternative distribution is proposed for ϕ: the $Beta(3, 5)$ distribution. The distribution of θ is unchanged. Explain how Monte Carlo integration can be used to estimate M, without sampling any new values of θ or re-evaluating the R function `profit`. If the original R analysis generated values $a_1, \ldots, a_{400}$, give a formula for the Monte Carlo estimate of P, corresponding to the new distribution of ϕ. ***(8 marks)***

[**Note:** The density of a $Beta(\alpha, \beta)$ random variable is

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\cdot, \cdot)$ is the Beta function.]

2 (continued)

(ii) 24 students taking a Maths course are split into three tutorial groups (A, B or C) each taught by a different tutor. There are 8 students in each group. At the end of the course, the exam marks for students in each class were recorded. In *R*, the exam scores and tutorial groups were stored in variables `exam` and `group` respectively

```
> group
 [1] A A A A A A A A B B B B B B B B C C C C C C C C
Levels: A B C
> exam
 [1]  79 70 91 75 84 77 85 75 74 89 67 83
 [13] 64 59 74 76 76 63 61 69 65 80 66 52
```

An observed test statistic is obtained and assigned to the variable `Obs` with the commands

```
> lm1 <- lm(exam ~ group)
> Obs <- summary(lm1)$fstatistic[1]
```

Two different methods are then used for testing the same hypothesis

Method I:

```
> # Method 1:
> N <- 10000
> T <- rep(0,N)
> for(i in 1:N) {
+   exam.new <- sample(exam, replace = FALSE)
+   lm.new <- lm(exam.new ~ group)
+   T[i] <- summary(lm.new)$fstatistic[1]
+ }
> mean(T >= Obs)
[1] 0.0229
```

Method II:

```
> # Method 2:
> N <- 10000
> T <- rep(0,N)
> for(i in 1:N) {
+   exam.new <- sample(exam, replace = TRUE)
+   lm.new <- lm(exam.new ~ group)
+   T[i] <- summary(lm.new)$fstatistic[1]
+ }
> mean(T >= Obs)
[1] 0.0221
```

(a) State the null hypothesis being tested and the alternative hypothesis. For each method, explain the procedure that is being used to implement the hypothesis test. ***(6 marks)***

2 (continued)

(b) What are the outcomes of the tests? ***(3 marks)***

(c) State the main assumption that is required for method I to be appropriate. ***(2 marks)***

(d) It is suggested that the p-value obtained using Method II will converge to the true p-value (for the observed data and choice of test statistic) as `N` is increased. Is this correct? Briefly justify your answer. ***(3 marks)***

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

$$X_i \sim \begin{cases} Poisson(\lambda) & \text{with probability } 1-\omega, \\ Poisson(\mu) & \text{with probability } \omega. \end{cases}$$

Let $\hat{\lambda}, \hat{\mu}$ and $\hat{\omega}$ denote the maximum likelihood estimates of λ, μ and ω given the observed data X only. The corresponding 'missing' variables $Y = \{Y_1, \ldots, Y_n\}$ are defined as follows:

$$Y_i = \begin{cases} 0 & \text{if } X_i \text{ is drawn from } Poisson(\lambda), \\ 1 & \text{if } X_i \text{ is drawn from } Poisson(\mu). \end{cases}$$

Define $\theta = (\omega, \mu, \lambda)$, with $\hat{\theta} = (\hat{\omega}, \hat{\mu}, \hat{\lambda})$.

(i) Using the fact that $P(X, Y|\theta) = P(Y|\theta)P(X|Y, \theta)$ for random variables X and Y, show that the log-likelihood

$$l(X, Y|\mu, \lambda, \omega) = \sum_{i=1}^{n} \Big[(1 - Y_i) \{\log(1-\omega) + X_i \log \lambda - \log X_i! - \lambda\} + Y_i \{\log(\omega) + X_i \log \mu - \log X_i! - \mu\} \Big].$$

(7 marks)

(ii) Show that the maximum likelihood estimates of ω, μ, λ given both X and Y are

$$\begin{aligned} \hat{\omega} &= \frac{\sum_{i=1}^{n} Y_i}{n}, \\ \hat{\lambda} &= \frac{\sum_{i=1}^{n} X_i(1-Y_i)}{\sum_{i=1}^{n}(1-Y_i)}, \\ \hat{\mu} &= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} Y_i}. \end{aligned}$$

(6 marks)

(iii) Use the factorisation theorem to show that sufficient statistics for ω, λ, μ given both X and Y are $\sum_{i=1}^{n} Y_i$, $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i Y_i$. ***(3 marks)***

(iv) Using Bayes' theorem, derive an expression for $P(Y_i = 1|X_i, \theta)$.

(5 marks)

3 (continued)

(v) Given X only, suppose the EM algorithm is to be used to obtain the maximum likelihood estimate $\hat{\theta}$ of θ. Let the current estimate of $\hat{\theta}$ be θ_{old}. By maximising

$$Q(\theta|\theta_{old}) = E[l(\theta; X, Y)|X, \theta = \theta_{old}],$$

show that the updated estimates of $\hat{\omega}, \hat{\mu}, \hat{\lambda}$ are

$$\begin{aligned}
\omega_{new} &= \frac{\sum_{i=1}^{n} p_i}{n}, \\
\lambda_{new} &= \frac{\sum_{i=1}^{n} X_i(1-p_i)}{\sum_{i=1}^{n}(1-p_i)}, \\
\mu_{new} &= \frac{\sum_{i=1}^{n} X_i p_i}{\sum_{i=1}^{n} p_i},
\end{aligned}$$

where p_i is your expression for $P(Y_i = 1|X_i, \theta = \theta_{old})$ in part (iv).

(9 marks)

4 (i) A set of $n+m$ patients in a hospital with bowel disease were given a new drug and the *remission time* (the time until their symptoms dissapeared) was observed in days. By the end of the trial not all of the patients had recovered, with some individuals still showing symptoms at the end of the study. The first n individuals were observed to fully recover with remission times $x_1, \ldots, x_n$. Patients $n+1, \ldots, n+m$ were still showing symptoms, with the i^{th} individual having been observed for c_i days, for $i = n+1, \ldots, n+m$.

The remission time X is modelled by a $Weibull(\lambda, k)$ distribution with density

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

(a) Show that, for an individual who is still showing symptoms after c_i days,

$$P(X > c_i) = e^{-(c_i/\lambda)^k} \quad c_i \geq 0.$$

(3 marks)

(b) Let $\mathbf{y} = (x_1, \ldots, x_n, c_{n+1}, \ldots, c_{n+m})$. Show that the log-likelihood for λ, k is given by

$$l(\lambda, k; \mathbf{y}) = n \log k - nk \log \lambda + (k-1) \sum_{i=1}^{m} \log x_i - \frac{1}{\lambda^k}\left(\sum_{i=1}^{n} x_i^k + \sum_{i=n+1}^{n+m} c_i^k\right).$$

(8 marks)

(c) Find the profile log-likelihood for k i.e.

$$l_p(k; \mathbf{y}) = \max_{\lambda} l(\lambda, k; \mathbf{y}).$$

(7 marks)

(ii) Suppose importance sampling, using a normal approximation as the importance density, is to be used to sample from a $Weibull(4, 6)$ distribution

$$f(x) = \begin{cases} \frac{6}{4}\left(\frac{x}{4}\right)^{5} e^{-(x/4)^6} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

By considering a Taylor series expansion of $\log f(x)$ about the mode of x, obtain good candidates for the mean and variance of the importance density. ***(12 marks)***

Note: The Taylor series expansion of a function $h(x)$ is

$$h(x) = h(a) + \frac{h'(a)}{1!}(x-a) + \frac{h''(a)}{2!}(x-a)^2 + \cdots.$$

to second order]

End of Question Paper